Novel expression sites and genetic diversity of *FoxP2* in Lake Malawi cichlids

A Thesis
Presented to
The Academic Faculty

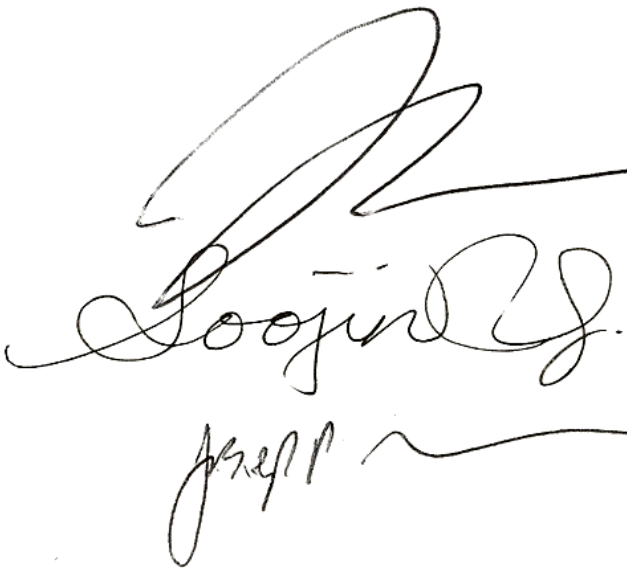
By

Michael Norsworthy

In Partial Fulfillment
Of the Requirements for the Degree
Bachelors of Science, Research Option in the
School of Biology of the College of Sciences

Georgia Institute of Technology
May 2011

Novel expression sites and genetic diversity of *FoxP2* in Lake Malawi cichlids



Approved by:

Dr. Todd Streelman, Advisor
School of Biology
Georgia Institute of Technology

Dr. Soojin Yi
School of Biology
Georgia Institute of Technology

Dr. Joseph Montoya
School of Biology
Georgia Institute of Technology

Date Approved: 5/3/11

I dedicate this work to my parents
Wayne and Patricia
And to my friends as well
Without whose
Love,
Friendship, and
Support
None of this work would have been possible.

And yes you lab people you,
That includes you too.

ACKNOWLEDGEMENTS

I am thankful for the mentorship, guidance, and advice of Eddie Loh, the immeasurable knowledge and expertise provided by Jon Sylvester, and the friendship and motivation by Todd Streelman. Their help not only brought this project to fruition, but furthered my development as a scientist and critical thinker.

TABLE OF CONTENTS

FIGURES AND TABLES

ABSTRACT

FoxP2 is a forkhead box-family transcription factor intricately linked to the development of learned motor skills, especially language in humans or vocalizations in other animals. It plays a prominent role in development and continues to be expressed in the adult. The gene has been thoroughly described in the brain, but transcripts have also been documented in areas such as in the lungs of mice (Shu et al, 2007). Mutations or deletions of *FoxP2* cause widespread defects in brain morphology, vocalization ability, balance and coordination, and even lung development. Broad conservation of *FoxP2*'s role in motor control and vocalization suggests the gene may play a similar role in vocally diverse Lake Malawi cichlids. Here, we probe cichlid *FoxP2* expression using *in-situ* hybridization throughout development, sequence and annotate the *FoxP2* gene across seven representative cichlid species, and search for polymorphisms. Expression in the brain, swimbladder, pharyngeal arches, and fins suggest *FoxP2* plays a role not only in sensory development and fine motor control, but also in the development of non-neural sound-producing structures as well as the development or innervation of skeletal muscle. Genetic analysis of *FoxP2* reveals polymorphisms which may be a source of species diversity in the development of any of the above functions, including vocalization. Two polymorphisms of note result in two amino acid changes near the DNA-binding domain of *FoxP2*. The natural phenotypic diversity of cichlid fishes makes them excellent candidates for further studies of the function and evolution of *FoxP2*.

INTRODUCTION

The conserved function of *FoxP2*

The forkhead box family of transcription factors contains dozens of regulatory molecules which play diverse and critical roles guiding development. Forkhead box P2 (*FoxP2*) has of late become an eminent member of the protein family following its implication in human speech disorders (Lai et al, 2003). First discovered in genealogical linkage studies of a British family termed "KE" with R553H mutations (Lai et al, 2003), orthologues of human *FOXP2* have since been found in organisms as diverse as the "mouse, songbird, frog, medaka, and zebrafish" (Bonkowsky et al, 2008), speaking to its deep ancestry and widespread conservation.

The conserved amino acid sequence of *FoxP2*

In humans afflicted with verbal dyspraxia, MacDermot et al (2005) documented a variety of mutations in *FOXP2* which correlate with the disease and do not occur in *FOXP2* sequences of control populations. As with the KE family, verbal dyspraxia patients often possessed R553H mutations. In addition, MacDermot et al found previously undocumented mutations such as a nonsense mutation (resulting in a severely truncated protein), small expansion of a polyglutamine stretch, and several substitutions including coding changes. The R553H and nonsense mutation alleles seem to be the most frequently recurring consequential perturbations of *FOXP2* (MacDermot et al, 2005). However, even though polyglutamine repeats are frequently implicated in neurodegenerative diseases, *FOXP2* polyglutamine repeats appear to be stable (Bruce & Margolis, 2002). Heterozygotes containing one normal *FOXP2* allele can also succumb to verbal dyspraxia, indicating correct dosing of the transcription factor is important (MacDermot et al, 2005).

Normal development of the cerebellum in mice is disrupted by gene knockout; cerebellar Purkinje cell layers are disrupted (Fujita et al, 2008), and overall size of the cerebellum is decreased (French et al, 2007). Mice homozygous for a nonsynonymous coding mutation (R552H, mirroring R553H in humans) in the DNA-binding motif of *FoxP2* experience broad and grave developmental deformations and die within a few weeks of birth, while heterozygotes survive but demonstrate significantly reduced motor-skill learning capacity (Groszer et al, 2008).

The amino acid sequence of *FoxP2* is very well-conserved: between zebrafish and humans, conservation of amino acids is higher than 80% (Shah et al, 2006). Incredibly, the sequence in mice and chimpanzees differs by only one amino acid (Enard et al, 2002), despite a distance of 84-121 million years since the last common ancestor to those lineages (Glazko et al, 2005). Additionally, deleterious mutations of amino acid sequence akin to those of the KE family or of the R552H mutation in mice experience strong negative selection. Amino acid sequence conservation throughout evolution readily indicates the gene plays a central, critical, and ditrillergic role as a regulatory molecule.

On the other hand, high linkage disequilibrium of *FOXP2* in humans indicates that it was recently the target of intense positive selection (Ptak et al, 2009). Human *FOXP2* differs from chimpanzee *FOXP2* by two amino acids, and from mouse *FoxP2* by three amino acids (Enard et al, 2002). The accumulation of two amino acid changes after the divergence of humans and chimpanzees strongly suggest direct functional consequences of those amino acid changes on vocalization and speech. Amino acid changes in *FoxP2* seem rare but potent in the courses of both disease and evolution.

Biochemical structure defines protein-protein interactions and downstream targets

Members of the FoxP subfamily contain a forkhead domain, a leucine zipper motif, a zinc finger, and a polyglutamine domain. The forkhead domain is common to all Fox proteins, directly binding DNA. However, unlike many Fox-family proteins, DNA binding activity of *FoxP1*, *FoxP2*, and *FoxP4* is contingent on dimerization (Li et al, 2004). Wang et al (2003) demonstrated that the leucine zipper in *FoxP1* is necessary and sufficient for *FoxP1* homodimerization. They also showed heterodimerization is possible between *FoxP1* and *FoxP3*. Meanwhile, the zinc finger could play a role in dimerization specificity (Wang et al, 2003). Further regulation may be provided by the polyglutamine domain, which is commonly a site of protein-protein interactions in transcription factors (Chamberlain et al, 1994).

Understanding of the role *FOXP2* plays in speech-related signaling pathways is incomplete, though many downstream targets of the gene have been identified. Vernes et al (2007) and Spiteri et al (2007) identified hundreds of *FOXP2*-targeted promoter binding sites in human tissue using chromatin immunoprecipitation microarrays (ChIP-on-Chip). Spiteri et al (2007) found that in most cases *FOXP2* acts as a transcriptional repressor. Binding targets include promoters belonging to functionally diverse genes such as those involved in cell communication, signal transduction, morphogenesis, maintenance of homeostasis, and ion transport (Spiteri et al, 2007). Vernes et al (2007) suggest target genes have roles in "modulating synaptic plasticity, neurodevelopment, neurotransmission, and axon guidance" (Vernes et al, 2007).

Vernes et al (2007) continued, suggesting that genes which have transcriptional responses to *FoxP2* often have at least two binding sequences within 100bp of each other in the promoter. This makes sense in light of the fact that *FoxP2* activity requires dimerization. *FoxP2* most

specifically binds (A)ATTTG(T), and is also capable of binding TATTTRT (generally a *FoxP1* site) and the Fox-family consensus sequence TRTTKRY. Various combinations of these consensus sequences may include two or more of the same consensus sequence, or a combination of different consensus sequences. Further, other transcription factor binding targets often co-occur, such as that for *LBP-1c*, indicating *FoxP2* dimers may be capable of acting in larger complexes (Vernes et al, 2007).

Many targets of *FoxP2* have been found to be transcriptionally dependent upon *FoxP1-FoxP2* heterodimers in the lung and gut (Shu et al, 2007). *FoxP1* expression has also been detected in the central nervous system during zebrafish development (Cheng et al, 2007), alluding *FoxP1-FoxP2* heterodimers could play some role in transcriptional control of nervous system development. In example, complex *FoxP1* and *FoxP2* co-expression hints that the developing striatum might witness heterodimers. An elegant pattern of expression is apparent in the straitum, with *FoxP1* expressed throughout the striatum and *FoxP2* localizing specifically to striosomes (Takahashi et al, 2003).

Broad spaciotemporal expression of *FoxP2* during development

FoxP2 is expressed early-on during development and is continually expressed throughout adult life. Expression profiles of *FoxP2* in vertebrates have consistently placed its expression domains throughout the brain, in such structures as the "telencephalon, basal ganglia, thalamus, tectum, tegmentum, cerebellum, and hindbrain" (Bonkowsky et al, 2008).

In 10 hours post-fertilization (hpf) zebrafish embryos, transcripts are already present throughout the embryo with a slight concentration towards the anterior (Bonkowsky & Chien, 2005). Large, densely concentrated expression domains develop first in the prosencephalon,

specifically the telencephalon by 24hpf and in the diencephalon by 36 hpf, with slight concentrations in the hindbrain. By 48hpf, expression dramatically increases in the midbrain at the tectum and the hindbrain at the cerebellum (Bonkowsky & Chien, 2005).

Takahashi et al (2003) performed a survey of *FoxP2* expression throughout development in the rat brain. At E13, the earliest stage tested, *FoxP2* expression is present in the ventral telencephalon. At E14, *FoxP2* transcripts are prominent in the ventral telencephalon, dorsal thalamus, zona limitans intrathalamica (ZLI), lateral ganglionic eminence, and the cortical plate. E16 rat embryos continue prominent expression in the dorsal thalamus, as well as expression in the striatum (fused from the lateral and medial ganglionic eminences) and preoptic area. Smaller expression zones are present in the piriform cortex, septum, and amygdala. Throughout E18, rat embryos continue prominent expression in the striatum, along with expression throughout the cortical plate and subventricular zone, and in the ventral cortex.

In post-natal P3 rats, Takahashi et al (2003) found that *FoxP2* adopts a mosaic expression pattern in the striatum, with patches of expression corresponding to striosomes. *FoxP2* continues to be expressed in striatal striosomes throughout the juvenile and adult, with other small points of expression at the shell of the nucleus accumbens and islands of Calleja (Takahashi et al, 2003). As *FoxP2* is expressed in the striatum of the basal ganglia throughout development (the basal ganglia develop from the ventral telencephalon) into the adult, *FoxP2*'s expression in the striatum seems of utmost importance to its function in mammalian brains.

Expression of *FoxP2* is prominent in the zebrafish adult (Shah et al, 2006), localizing mainly to the periventicular gray zone on the subcortical side of the optic tectum. Transcripts are also present in the adult preoptic area, ventral telencephalon, hypothalamus, and the caudal lobes of the cerebellum (equivalent to the mammalian vestibulocerebellum) (Shah et al, 2006).

Regulation of *FoxP2* during development and sensorimotor learning

Clearly, *FoxP2* is expressed throughout development and into adulthood, though the regulatory mechanisms that control its expression are less clear. Only a handful of transcription factors have been identified which control *FoxP2* expression. *Lef1*, activated through Wnt signaling, appears necessary for *FoxP2* expression in some neural tissues during development (Bonkowsky et al, 2008).

Interestingly, in juvenile and adult zebra finches, *FoxP2* levels are actively regulated during song learning (Teramitsu et al, 2010). Most song learning occurs in juveniles, though adults also practice and refine their songs periodically. *FoxP2* is transiently suppressed for several hours following song practice, proportional to the duration of practice. Specifically, *FoxP2* expression is suppressed in Area X of the striatum.

Juvenile finches singing the morning after practicing showed higher variability in song, allowing them to explore and fine-tune their song; this process is beneficial for refining a song to its ultimate adult form. Direct suppression of *FoxP2* with interfering RNA duplicated the increased song variability normally caused by practicing.

Teramitsu et al (2010) propose that *FoxP2* "could function as a 'plasticity gate'", allowing variability of synapses while low and promoting stability of synapses while high in level. As juveniles practice more than adults, *FoxP2* is more often downregulated in juveniles, possibly allowing greater neuronal plasticity while neuronal circuits are being developed for song learning. This makes sense in light of the fact that *FoxP2* generally acts as a transcriptional repressor, targeting genes such as those involved in Notch and Wnt signaling, axon guidance, neurotransmitter release, and other nervous functions (Vernes et al, 2007).

A confounding level of complexity still exists for unraveling how *FoxP2* is regulated. An unknown number transcription factors regulate *FoxP2* expression, outside of those already identified such as *Lef1* (Bonkowsky et al, 2008). The picture is further muddled by multiple spliceoforms (Bruce & Margolis, 2002). Further, at least 4 transcription entry sites are present in humans, each giving rise to different pre-mRNAs which may be preferentially transcribed according to cell type (Schroeder & Myers, 2008).

Cichlids as a model for investigation

Cichlid fishes of the East African Great Lakes collectively represent a remarkable resource for evolutionary studies, exhibiting widely-sweeping diversity of behavioral and morphological phenotypes despite differing very little on a molecular basis (Klein et al, 1993) and being defined by only 10 million years of divergence (Turner et al, 2001; Kocher, 2004). Sexual selection is a primary contributor to speciating divergence, maintaining mating isolation in sympatric populations of East African cichlids (Stauffer et al, 1995). Conspecific mating disfavors species hybrids which may otherwise be fecund and well-adapted for natural selection (Van Der Sluijs et al, 2008).

Vocal diversity of Lake Malawi species

Vocalization during courtship is a well-documented behavior for assortative mating and sexual selection (Amorim et al, 2004; Fryer & Iles, 1972; Lobel, 1998). Ripley & Lobel (2004) contend that the ability for *Tramitichromis intermedius* males to produce sound during courtship displays is considered by females assessing "mate quality". Lab-raised and wild *T. intermedius* show no difference in vocalization, indicating courtship rituals are heritable (Ripley & Lobel, 2004). Vocal behavior is a genetically based trait that sexual selection can act on.

Courtship vocalizations have common structure and several properties which can be quantified (Lobel 1998, Amorim et al 2004). Each call consists of a number of pulses and periods of interpulse silence. Most obviously, the rhematonic number of pulses per call can be measured. Overall call duration as well as the duration of pulses and interpulse periods can each be measured in units of time. Further, each pulse's frequency can be measured in Hertz.

Several species of Lake Malawi cichlids have been shown to have significantly different patterns of sound production. Lobel (1998) recorded and compared *Mchenga conophorous* (previously *Copadichromus conophorus*) vocalizations to those of *Tramitichromis* cf. *intermedius* in the wild. He found significant differences in pulse rate and pulse duration, and no significant differences in pulses per call, call duration, or interpulse interval.

Amorim et al (2004) recorded sounds produced by sympatric populations of closely related species, *Pseudotropheus zebra*, *P. callainos*, and *P.* 'zebra gold'. Peak frequency differed significantly between *P. callainos* and *P.* 'zebra gold'. *P. callainos* also had significantly longer pulse duration than *P. zebra*. Importantly, Amorim et al showed that peak frequency can differ significantly between species, even when considering differences in body size. A linear fit of *P. callainos* peak frequency versus body standard length revealed a higher frequency per unit body

length than a linear fit for *P.* 'zebra gold'. Thus, the frequency of call is determined though factors other than simple allometrical scaling with body size.

Possible mechanisms of sound production in cichlids

A leading hypothesis holds that pharyngeal jaws are primarily responsible for sound production in cichlid fish (Lanzing et al, 1974; Lobel, 2001). Rice and Lobel (2002) provide additional evidence for this hypothesis. In the sexually dimorphic *Tramitichromis intermedius,* pharyngeal jaw muscles in males display significantly higher β-oxidative and anaerobic ability than female jaw musculature. As the *T. intermedius* male is vocal while the female is not, the physiology of the pharyngeal jaw appears important to sound production.

Longrie et al (2009) have proposed another possible mechanism of sound production in cichlids involving an interaction between hypaxial musculature and the rib cage and swimbladder. These hypotheses are not necessarily mutually exclusive, and may correctly model different lineages of fish. Longrie et al (2009) suggest that in *Oreochromis niloticus,* pharyngeal jaws do not play a role in sound production. They reason that *O. niloticus* jaws produce sound at 2,700 Hz during chewing, while sounds during territory defense have an average frequency less than 200 Hz. Further, the fish may be deaf above 2,000 Hz, and physical obstruction of pharyngeal jaws does not inhibit sound production during electrostimulation.

The authors continued, showing that *Oreochromis niloticus* has an anatomy that is potentially well-adapted to making sound via their proposed mechanism. Part of the axial musculature, which the authors term the vesica longitudinalis, is a symmetrically paired structure along the anterior-posterior axis, at both grooves shaped by the junction of where the swimbladder and peritoneal cavity meet. *O. niloticus* initiates sound production by retracting the

pectoral girdle posteriorly and the pterygiophore anteriorly, resulting in compression of the rib cage and swimbladder.

Meanwhile, *Cyphotilapia frontosa* has a morphology with less-obvious vocalization capability: it has no grooves between the swimbladder and peritoneal cavity and no well-defined functional equivalent to the vesica longitudinalis muscle. In addition to anatomical observations, the authors shocked specimens of *C. frontosa* and *O. niloticus* with electrostimulation. *O. niloticus* produced sound involuntarily during shocking, while *C. frontosa* produced no detectable sound (Longrie et al, 2009). A groove and vesica longitudinalis, in coordination with the swim bladder, appear to be important to sound production in some cichlid species.

Problems remaining in the field

Comprehensive understanding of the genetic and biochemical interactions of *FoxP2*

In spite of some study of *FoxP2* expression in the brain of various species, its biochemical role in the larger problem of development is less well-defined. Its interactions with FoxP and other transcription factors present biochemical complexity in regulating downstream genes. Though many downstream targets of *FOXP2* have been identified (Spiteri et al, 2007; Vernes et al, 2007), there lies a poorly characterized elegance of the full genomic regulatory network in which *FoxP2* interacts. As Konopka et al (2009) show using ChIP-on-Chip to compare human and chimpanzee *FOXP2* promoter occupancy, *FOXP2*'s regulatory action is extremely broad and complex, and extremely contingent on its amino acid sequence.

Though some upstream regulators of *FoxP2* have been identified such as *Lef1*, the view of mechanisms that regulate *FoxP2* is incomplete (Bonkowsky et al, 2008). Multiple spliceoforms and multiple transcription entry sites suggest massive potential for regulating the

translated amino acid sequence in a tissue-specific fashion (Bruce & Margolis, 2002; Schroeder & Myers, 2008). Multiple isoforms present an even greater complexity, where the isoform may define protein-protein interactions such as between *FoxP2* and *FoxP1*. All these regulatory mechanisms are equally as important as understanding the downstream targets of *FoxP2* regulation in order to meet the goal of placing *FoxP2* into a comprehensive regulatory network.

Full characterization of *FoxP2* expression

Though most characterization of *FoxP2* has focused on the brain, *FoxP2* is also expressed outside the central nervous system. Expression has been detected in the lungs and the gut, indicating that expression is not limited to the central nervous system or even to one tissue layer (Shu et al, 2007). *FoxP2*'s massive potential for expression and complex interactions necessitates a more careful and comprehensive survey of its expression throughout the process of development. Most importantly, *FoxP2* may guide the development of speech-related structures besides the brain, such as the lungs, airways, larynx, pharynx, and mouth in mammals.

Given that the swim bladder (SWB) may play an important role in fish vocalizations, and that physiology may directly determine vocalization ability such as between *C. frontosa* and *O. niloticus*, the evidence pose a fascinating question: what factors are responsible for the development of these structures, and how do they differ between vocally discrete species? The implications are beyond academic in nature. In fishes, the swim bladder (SWB) has traditionally been viewed as a homologous structure to the terrestrial vertebrate lung, based on its function and morphology (Graham, 1997; Perry et al, 2001). Winata et al (2008) provide conclusive developmental and molecular evidence for SWB homology to the lung. Since *FoxP2* expression is present in the mouse lung (Shu et al, 2007), transcripts also may be present in the SWB. Given

that the SWB may play a role in cichlid vocalization, *FoxP2* expression in the SWB may have phenotypic consequences for vocalization anatomy outside the nervous system.

Similarly, the pharyngeal jaws in fishes have homology to the mammalian larynx in that both develop from pharyngeal arches. The larynx, descended in humans relative to other great apes, may explain some differences of speech ability (Lieberman, 2007; Boë, 2007). Evoking potent similarities to studies of the larynx, Rice and Lobel (2002) demonstrated that the physiology of pharyngeal jaws can define divergent vocalization ability between sexes of *T. intermedius*. Could *FoxP2* play a role in guiding development of pharyngeal structures in a species-specific manner? Could *FOXP2* play a role in larynx formation in humans? Could the amino acid changes in *FOXP2* in humans be responsible for the descent of larynx not present in other extant great apes? Better characterizations of *FoxP2* expression in cichlids may shed light on these questions and contribute to their eventual answers.

Mechanisms of conspecific mate choice of Lake Malawi cichlids

On the other hand, despite some forays into understanding the mechanisms behind cichlid sound production, there still exists serious lack of knowledge into how this important factor of assortative mating operates. As East African cichlids become more popular in comparative genomics and evolutionary development, understanding the basic mechanisms behind the causes of their divergence is key. Even less understanding exists regarding the biomolecular chemistry behind the factors on which assortative mating operates.

OBJECTIVES

FoxP2 may play a yet uncharacterized role in helping to pattern speech and motor-related structures outside the brain. As the use of the pharyngeal jaw (Rice and Lobel, 2002), or alternately the swim bladder and vesica longitudinalis (Longrie et al, 2009), may be responsible for sound production, one could reasonably suppose that *FoxP2* might be expressed during development of those structures, perhaps placing it in the pharyngeal arches and swim bladder during development.

We propose that *FoxP2* could be involved in the patterning, morphogenesis, or innervation and control of vocal organs as well as skeletal muscles throughout the body, particularly in muscles which require fine motor control. Using *in-situ* hybridization, we are determined to provide a comprehensive view of the expression of *FoxP2* throughout development. In addition, we will attempt to collect expression data across species and determine if differences in expression may correlate with species-specific vocal properties.

We further propose that the behavioral diversity, including vocal diversity, of Lake Malawi cichlids may be in part explained on a molecular level by differential regulation, expression, or protein action of *FoxP2*. We will sequence the entire *FoxP2* gene across multiple cichlid species, find conserved regions relative to outgroups of other fish and mammals, and survey genetic diversity across cichlid species in the gene. In conserved regions, we will attempt to predict binding sites *in silico* of regulatory factors including and in addition to *Lef1*. Polymorphisms at putative regulatory sites will be suspects for future screening and research.

Specific Aim 1

Clone and characterize *FoxP2* in cichlids

Before further study may take place, *FoxP2* must be cloned in cichlids. A basic cichlid *FoxP2* mRNA sequence is a necessary foundation on which subsequent study must be built. With that knowledge in hand, it becomes possible to probe for mRNA transcripts in expression studies.

To gain an appreciation for expression of the gene, *in-situ* hybridization will be performed on at most two to three species across several representative timepoints of development. In particular, the developmental stages of somatogenesis, neurulation, and later stages will be closely monitored for *FoxP2* expression throughout the embryo.

Initial characterization of expression is intended to be spatially comprehensive and representative of the major timepoints in development. Key expression domains are expected to be found throughout the brain, in the pharyngeal arches and developing pharynx, and the swim bladder.

Specific Aim 2

Characterize differential *FoxP2* expression between species

From initial expression data, higher-resolution monitoring will be performed with the intent of finding divergent expression between species. *In-situ* hybridization will be expanded to at least three species to compare spaciotemporal expression between them. At a given time point, large sample sizes can allow comparisons of spatial expression between species. Meanwhile, data collection at high temporal resolution (sacrifice of embryos every 2-4 hours throughout development) will permit comparisons of expression timing across species. Divergence may appear at any time point and any expression site. Differential brain expression would imply

differences in motor control. Differential expression in the pharyngeal arches and developing pharyngeal jaws, or at the swim bladder, would imply physiological differences directly affecting vocal structures.

Specific Aim 3

Investigate the molecular evolution of *FoxP2* among Lake Malawi cichlids

Sequence data of *FoxP2* genomic DNA will be obtained for several cichlid species in order to identify single nucleotide polymorphisms (SNPs). Any SNPs located will be genotyped to assay the amount of allele fixation between divergent cichlid populations. Highly fixed SNPs will be of particular interest for further study of their potential functional consequences of expression regulation or protein structure.

Impact

Investigating the divergence in sequence or expression *FoxP2* in Lake Malawi cichlids

We expect that this project will make key contributions to the understanding of how *FoxP2* is regulated, lending further illumination to this prominent yet only recently scrutinized transcription factor and its role in the development of neural and non-neural tissues involved in producing vocalizations.

This investigation is at the crossroads of molecular biology and evolutionary development, and represents one of the first attempts to provide a molecular-based explanation for one of the most intriguing drivers of cichlid evolution. The causes of divergence between cichlid vocalizations may provide direct insight into conserved regulatory mechanisms of genomic regulatory networks involving *FoxP2*. By exploring how these mechanisms differ

between divergent cichlids, crucial links can be made in surmising the real functional relationships behind the relevant genetic interactions. This research should help construct a more complete picture into how this widely conserved and broadly expressed transcription factor plays its part in the wider puzzle of development.

By necessity from *FoxP2*'s highly conserved function, lessons learned here will apply directly to research on the mechanisms behind speech and sensorimotor-related development and behavior, and may prove valuable to understanding the causes of developmental conditions such as verbal dyspaxia. Further, as *FoxP2* is believed to be an original driver of human evolution, research into this molecule can provide a small but important contribution in understanding the overall large and complex picture of our ancestors' history.

METHODS AND MATERIALS

Lake Malawi fishes

Sustenance of fish

All live fish were cared for in accordance to Institutional Animal Care and Use Committee guidelines. Fish were fed algae flakes once daily. Fish tanks were continually supplied with filtered and UV-sterilized water. Oxygen concentration in the water was maintained at full saturation by bubbling filtered pressurized air into each tank.

Selection of model species

In gathering expression data, *Metriaclima zebra* embryos were preferentially selected for *in-situ* hybridization, though at times brood availability made selecting *Labeotropheus fuelleborni*, *Mchenga conophorus*, or *Aulonocara jacobfreibergi* embryos more practical.

The following model species were chosen for genetic analysis: (A) *Copadichromis eucinostomus*, (B) *Cynotilapia afra*, (C) *Protomelas taeniolatus*, (D) *Labeotropheus fuelleborni*, (F) *Metriaclima zebra*, (G) *Tramitichromis brevis*, and (H) *Mchenga conophorus*. *Mbuna* species are *C. afra, L. fuelleborni*, and *M. zebra*. Non-*mbuna* species are *C. eucinostomus, P. taeniolatus, T. brevis*, and *M. conophorus*.

Isolation of DNA

Isolation of genomic DNA

Anal fin clippings were taken and total genomic content extracted using a Qiagen DNeasy Blood & Tissue Kit according to manufacturer's protocol. Some genomic DNAs were amplified using a GenomiPhi kit to preserve their stock.

Isolation of total RNA from embryos

Total embryonic RNA content was purified from embryos using a Qiagen RNeasy Mini Kit according to manufacturer's protocol.

Obtainment of *FoxP2* cDNA

Reverse transcription of isolated total RNA was performed using a Clontech SMART MMLV Reverse Transcriptase kit according to manufacturer's protocol.

Genetic approaches

Retrieval of genetic and genomic information for cichlid comparisons to orthologues

Coding sequences, protein sequences, full transcripts, and genomic assemblies containing the *FoxP2* gene were downloaded for Tetraodon, Fugu, Stickleback, Zebrafish, Medaka, Mouse, Chimpanzee, and Human from the Ensemble database (European Bioinformatics Institute).

Retrieval of Tilapia sequence to aid primer design

In the absence of an assembled cichlid genome, a draft contig of unannotated Tilapia sequence mapped to a *Gasterosteus aculeatus* scaffold was downloaded from bouillabase.org (NODE_2193628_length_398943). Exons from *FoxP2* orthologues in Stickleback and Medaka were individually aligned to the Tilapia sequence using a BLASTn local alignment tool. A non-comprehensive Tilapia mRNA sequence was assembled from exons located in the unannotated Tilapia sequence. Due to the minimal evolutionary distance between Tilapia and Lake Malawi cichlids, sequence homology was expected to be high.

Primer design

Using Primer3, primers were designed off Tilapia sequence for targeting either transcript or genomic sequences. Reactions were generally designed to yield 600-900 nucleotide products, including 100-200 nucleotide overlaps of consecutive amplicons. Primers were designed in two sets. (1) Primers were designed to amplify only the transcript sequence, and were limited to the domains of predicted exons. (2) Primers targeting the entire genomic sequence were designed without any preference for intronic and non-transcribed sequences in addition to exons. The genomic set of primers was designed to target sequence along the entire set of exons and introns, along with about 5-10 kb of sequence upstream of translation start and 5-10 kb downstream of translation end. These sets of primers are detailed in Supplementary Tables S1 and S2 respectively.

Polymerase Chain Reaction (PCR)

Amplification was carried out using 0.5uL of each designed primer at 10uM, 1 uL of cDNA or genomic DNA at >20nM, and 2X Go-Taq Hot Start PCR Reagent diluted to 1X with DEPC-H_2O. Reactions were programmed on a thermocycler. Initial denaturing of double-stranded DNA ran 3 minutes at 94.0 °C, followed by 35 cycles of (1) 30 seconds denaturing at 94.0 °C, (2) 45 seconds for primer annealing to single-stranded DNA at 54.0 °C, and (3) novel strand extension by Taq polymerase at 72.0 °C for 1 minute 30 seconds, with a final extension time of 7 minutes at 72.0 °C.

Sequencing

High-throughput sequencing of genomic PCR amplicons was performed by High-Throughput Sequencing Solutions at the University of Washington, Department of Genome Sciences. In compliance with their protocols, samples were submitted in a paired 96-well plate format, where one 96-well plate contained 10uL of template DNA and the matching 96-well plate contained 10uL of one primer at 10uM concentration. Chromatograms and base calls were downloaded, and quality scores were calculated by Sequencher.

Towards a draft *FoxP2* coding sequence (CDS)

Fragments of cichlid exonic sequence were cloned off of *Metriaclima zebra* cDNA by PCR using primers designed on Tilapia exons (as described above), then sequenced. A draft 1x coverage of cichlid *FoxP2* coding sequence was assembled in Sequencher. The proposed sequence was checked using tBLASTn against the translated nucleotide database of GenBank. Significant homology was found in *FoxP2* of other taxonomic lineages including *Homo sapiens*, with much lower sequence similarity versus other *foxp* members in any lineage.

As this draft CDS was constructed with approximately 1x coverage, its sequencing quality required caution when looking at its sequence on the level of individual nucleotides. In particular, its read quality near the 5' and 3' extremities approached high rates of erroneous base calls. Nevertheless, the majority of base calls provided an excellent initial assembly to anchor subsequent investigation (such as probe design, discussed below), leaving refinement of the sequence to be provided by later sequencing of *FoxP2*'s genomic sequence and exon annotation.

Assembly of the *FoxP2* genomic region in seven species and implicit 7x consensus coverage

Genomic-designed primers (137 pairs successful) were used in PCR reactions for each of 7 species. Reactions were checked for success on a 1% agarose gel then sequenced.

Homologous amplicons from each single individual were assembled using Sequencher into an alignment of equivalent sequences from different individuals. Then, Sequencher was used to join overlapping tandem alignments end-to-end into a continuous sequence. A consensus sequence was derived from the agreement of aligned sequences. This consensus represents the "average" Lake Malawi cichlid *FoxP2* gene with approximately 7x coverage, though some areas are less well-covered.

The consensus was exported as a single sequence, and imported into seven separate Sequencher files corresponding to sequencing reads for each of seven species. For each species, all sequencing reads were quality-trimmed (from each end, trim: no more than 25%, until 12 bases contain ≤1 bases with quality <20). Gaps in an individual's sequence were marked with dashes ("-"). Ambiguities were marked with Ns. Each of these seven individual assemblies were exported in FASTA format from Sequencher. Assemblies were all approximately 75 kb but none were the same length as the consensus due to small insertions and deletions or slight misalignments in Sequencher. Assemblies were collected and aligned in ClustalW such that each assembly became the same length, making subsequent analysis easier.

The coding domain of cichlid *FoxP2* in seven species and 7x consensus coverage

The ~7x coverage genomic consensus sequence was locally aligned with the 1x draft coding domain sequence using the Blastn algorithm on NCBI's bl2seq. The draft CDS fully mapped onto the genomic sequence in 16 significant high-scoring pairs, representing 16 rough

exons. Exon assignments were manually inspected for splice signals and frame of translation, adjusted as appropriate, and assembled into a proposed revised CDS. This putative CDS was translated one final time for confirmation of correct frame, and confirmed to be in frame through similarity to other *FoxP2* proteins. Past this point, the putative CDS accumulated enough translational and genomic evidence to be considered a true representation of the cichlid *FoxP2* coding domain (Fig. 7).

The consensus coding domain was then locally aligned (one exon at a time) against each of the seven individual 1x genomic assemblies. Retrieved hits representing exact exons were assembled into individual CDSes. Individual coding domains were therefore 1x coverage, with some gaps and low-quality reads in some areas of the data set for a given individual.

Discovery of putative single-nucleotide polymorphisms (SNPs)

Each base position in genomic and coding-domain sequences was analyzed for disagreements ignoring gaps and Ns. Putative SNPs were identified when two or more types of non-gap, non-ambiguous bases were present at the same position in different species.

Expression profiling

Probe design

Within an area of strong read quality within the draft transcript (before the final CDS was completed), we split the assembled sequence into fragments matching our predicted homologous Tilapia exons. Moving exon by exon, we checked specificity of sequence to *FoxP2* compared to paralogous genes like *FoxP1* and selected a ~500 nt sequence which contained exons specific to *FoxP2*. Primers chosen for isolation were left: 5'-TGTCAGTGGCCATGATGAGT-3' and right: 5'-CTGTGTTTGATGCCGTTGTC-3'.

The selected sequence was amplified from cDNA in PCR, ligated into pGEM-T Easy Vector according to manufacturer's instructions, and incubated at 4 °C overnight. Ligation product (3 uL) was transformed into JM109 competent cells on LB/Amp/IPTG/X-Gal media. After over-night incubation at 37 °C, white colonies were saved on fresh LB/Amp media and verified for the insert with PCR and visualization on a 1% agarose gel. One colony testing positive for the insert was selected and grown up in 150 mL LB/Amp overnight at 37.0 °C.

Plasmid was extracted using a QIAGEN Maxi-prep kit with the manufacturer's protocol. The extracted insert was sequenced to 6x coverage (Figure 1). The plasmid was linearized with 2uL 10X Buffer, 11uL DEPC-H_2O, and 2uL SpeI restriction enzyme (checked to cut the Sp6 site but not our insert) for 2 hr at 37.0 °C to digest.

Digest was cleaned and resuspended in nuclease-free H_2O. The insert was reverse transcribed into antisense with T7 polymerase, precipitated in EtOH, resuspended in nuclease-free H_2O, and kept at -20 °C.

Figure 1: Sequence of the insert, given as the coding strand.

```
TGTCAGTGGC CATGATGAGT CCCCAGGTGA TGACGCCGCA GCAGATGCAG CAGATCCTCC
AGCAGCAGGT GCTTTCCCCC CAGCAGCTCC AGGCCCTGCT CCAGCAGCAG CAGGCTGTAA
TGCTGCAGCA GCAACACCTG CAGGAGTTTT ATAAGAAACA ACAGGAACAG CTTCATCTGC
AGCTTCTGCA GCAGCAACAC CCTGGCAAGC AGGCTAAAGA GCAACAGCAG CAGCAGCAGC
AGCAGCAGCT GGCCGCCCAG CAGCTCGTCT TCCAGCAGCA GCTCCTGCAG ATGCAGCAGC
TCCAACAGCA GCAGCACCTG CTCAACATGC AGCGGCAGGG CCTGCTCACG CTGCCTGGTC
CCGCTCCAGG CCAAGCCGCC CTGCCCGGAC AGACCCTACC CCCACCAGCT GGATTGAGTC
CCGCAGAGCT CCAGCAGTTG TGGAAGGACG TTACCGGAGG CGGCGGTCAC GGAATGGAGG
ACAACGGCAT CAAACACAG
```

In-situ hybridization (ISH)

In-situ hybridization was performed as described in Sylvester et al (2010) and adjusted as appropriate for cichlid embryos at various developmental stages. PFA (4%) was used to fix embryos for 2 days. Embryos were then rinsed twice and washed (10 min each) three times in PBST, then successively dehydrated in 3 increasing graduations (25%, 50%, 75%) of MeOH in PBST and 2 washes of 100% MeOH (10 min each). Dehydrated embryos were incubated at -20 °C overnight, then rehydrated into PBST through 3 decreasing increments (75%, 50%, 25%) of MeOH in PBST (5 min each), culminating in 2 washes of 100% PBST (5 min each).

Embryos were dechorionated if less than 5 days old. If embryos were greater than 3 days old, digestion with Proteinase K was performed. A wash in prehybridization solution was given until embryos lost buoyancy. Then, embryos were incubated in fresh prehybridization solution at 70 °C for 2.5 hr. Prehybridization solution was replaced with hybridization solution containing 20 uL probe per mL solution. Embryos were incubated at 70 °C overnight for hybridization.

Embryos, still at 70 °C, were then washed in prehybridization solution twice (5 min each), 25% prehybridization solution/75% 2X SSC (5 min), 2X SSC (10 min), then three times in 0.2X SSC (30 min each). Continuing at 20 °C, they were briefly rinsed twice in MABT.

Subsequently, they were incubated in blocking solution for 2.5 hr at 20 °C with shaking, then introduced to AP Fragments Anti-DIG antibody at 1:3000 dilution in fresh blocking solution. Incubation proceeded overnight at 4 °C with shaking.

Blocking solution was removed with two MABT washes at 20 °C (5 min each). Then, 5 washes of TST (1 hr each) were administered. Embryos were then washed twice in NTMT (5 min each), then introduced to NTMT with NBT/BCIP (20uL per mL NTMT) in the absence of ambient light. After running 3 hours, two washes of PBS stopped the coloring reaction. Embryos were fixed once again in 4% PFA (2 hours), washed in PBS once, and kept in PBS at 4 °C.

Whole-mount visualization

After ISH, embryos were directly visualized in the dorsal or lateral view.

Sectioning

After ISH, embryos were suspended and blocked in 900uL egg with 100uL glutaraldehyde and allowed to set. Solidified egg was fixed for at least 4 hours in 4% PFA. Blocks were kept wet in PBS. Embryos were sectioned using a vibrating microtome in the sagittal, dorsal, or coronal planes with 20, 25, or 30um sections.

RESULTS

Expression of *FoxP2* throughout development

Expression in the 3-day *M. zebra* embryo

At the time of this writing, 3 day embryos are the earliest-stage cichlids included in our still-expanding dataset. Even at this early timepoint, the expression of *FoxP2* was well-established in the anterior portion of the fish, particularly in the brain and eyes, and extended down the neural tube.

In the dorsal view, the most noticeable expression is in the midbrain, clearly visible even in whole-mount (Fig. 2A). Expression is also visible in the developing eyes, particularly at their posterior boundaries. Also of note in the dorsal view are a left-and-right pair of expression zones at the anterior-most extent of the fish, and a medial band of expression running posterior throughout the fish.

The sagittal view (Fig. 2B) reveals that in the forebrain, three foci of expression approximately anterior to the prospective ZLI are visible, the anterior-most of which includes one member of the left-right pair mentioned above (Fig. 2A). Further posterior, a dark zone of expression extending from the ventral-most to dorsal-most part of the fish is visible. This zone corresponds to the large zone visible in the dorsal view (Fig. 2A).

Coronal sections reveal additional details about *FoxP2* expression in 3 day embryos. The plane gives us an additional view of the small paired points of expression seen in dorsal and sagittal orientations (Fig. 2C). A particularly dark area of expression is visible immediately between the eyes. Additionally, the coronal view reveals a two-layered expression pattern of *FoxP2* corresponding to an outer and middle layer of the developing eyes (Fig. 2D). Posterior to the eyes, further detail is provided about the dark zone of expression visible in dorsal and sagittal

views. The zone, which in sagittal view (Fig. 2B) looks imperfectly defined, takes on the appearance of three bands at different points along the dorsal-ventral axis (Fig. 2E). Past the brain, *FoxP2* expression continues down the neural tube along its most ventral point (Fig. 2F).

Expression in 5-day (hatching) 5 *M. zebra* embryos

FoxP2 continued its prolific expression in the central nervous system. Most anterior, parts of the developing telencephalon stained positive for the transcript. Here, expression was visible as small foci in the pallium and large domains throughout the subpallium (Fig. 3B). The olfactory bulbs showed no signs of expression.

In the diencephalon, expression was present as a banded pattern in the dorsal thalamus and the ventral thalamus (Fig. 3B, 3C). Dark expression domains were visible in the pretectum, distal from the midline and proximal to the eyes (Fig. 3C). On the other hand, transcript was not detected in the hypothalamus or most of the posterior tuberculum, both members of the basal diencephalon (Fig. 3B, 3C, 3D).

The mesencephalon also showed widespread *FoxP2* expression and possessed the largest contiguous volume of *FoxP2* expression throughout the fish in the optic tectum. The optic tectum, extending as a large, paired shield-shaped structure on the dorsal side of the brain, displayed a large swath of *FoxP2* expression throughout its core (Fig. 3A, 3B, 3C, 3D). The tegmentum, immediately ventral to the optic tectum, also stained positive for *FoxP2* transcript.

Expression was also present in the hindbrain (Fig. 3B). The prospective cerebellum (rhombic lip) expressed *FoxP2* in its central area. Additionally, prominent domains in the medulla oblongata stained for hybridized transcript. Past the hindbrain, expression continued down the spinal cord (Fig. 3A).

Notably, *FoxP2* messenger RNA was detected in many structures outside of the CNS. The eyes featured expression once again in two layers, in their ganglion cell layer and inner nuclei layer (Fig. 3C). The pharyngeal arches were also positive for *FoxP2* (Fig. 3F).

Expression in 7-day (early larval) *M. zebra*

In 7-day embryos, *FoxP2* continued to be expressed in the brain without any visible signs of abatement. The same brain structures as in hatching embryos expressed *FoxP2*, though as the structures continued to become better elaborated, expression patterns became more well-defined.

The subpallium maintained its strong expression, and the pallium continued weaker but still visible expression. The domain of expression was contiguous between the subpallium and pallium, rooted in the subpallium and extending in tendrils to the medial and the lateral pallium (Fig. 4C). The dorsal and ventral thalamus continued to express *FoxP2* as well, but this pattern became more well-defined, manifested as lateral bands throughout these regions (Fig. 4D). Anterior parts of the posterior tuburculum also appeared to transcribe the gene (Fig. 4E).

The pretectum led expression domains into the midbrain's optic tectum (Fig. 4D). Transcripts were detected throughout the optic tectum at this stage, throughout its core and extending almost to its boundary with the rhombic lip (Fig. 4D, 4E, 4F). Transcription continued uninterrupted in the hindbrain as well, present in the rhombic lip and medulla oblongata (Fig. 4F). The neural tube continued to express *FoxP2*. This pattern was more complex than in 3 day embryos, transitioning from a medial point along the ventral side of the neural tube to several paired bands throughout the neural tube (Fig. 4G, 4H; compare to Fig. 2F). In addition, this banded pattern seemed to be variable in strength along the anterior-posterior axis. Stronger areas on both the left side and right side of the neural tube were generally paired (Fig. 4I).

Outside the central nervous system, the pharyngeal arches seemed to increase their production of *FoxP2* transcript (Fig. 4B, 4D, 4E, 4F) to more readily visible levels (see Fig. 6 for additional evidence of pharyngeal arch expression). Meanwhile, the pectoral fin buds had grown into visible structures, and the gut began transitioning to a more coiled form evidenced by its asymmetry in some sections (compare 4G, 4H). The mesenchyme of developing pectoral fins stained positive for transcript (Fig. 4A, 4G), as did the developing caudal fin (Fig. 4A). The gut displayed some transcription around its periphery (Fig. 4G, 4H). Interestingly, an intense domain of expression materialized in an area bounded ventrally by the gut, laterally by mesodermal tissues, and dorsally by the dorsal aorta and notochord (Fig. 4H). This area, slightly posterior to the fins, will shortly be home to the swimbladder.

Expression in late larval *M. conophorus* embryos

Cichlids later in development displayed a noticeable shift in their expression of *FoxP2*. The gene was still expressed in the brain and central nervous system, but its pattern changed visibly, weakening in previously strong areas of expression such as the cerebellum, though remaining strong in the optic tectum and forebrain structures (Fig. 5A). On the other hand, expression increased further in complexity in the spinal cord (previously the neural tube), becoming pronounced in discrete ganglia (Fig. 5B; compare to 4H, 3E, 2F).

The fins also displayed increasingly complex patterns of transcription. The pectoral fins, the first fins to develop (Fig. 4A, 4G) continued expression in their mesenchyme, while developing bony tissue did not stain positive (Fig. 5E). The later-developing unpaired dorsal and anal fins showed an alternating pattern of spline, expression, spline, expression, and spline progressing from the anterior to posterior direction (Fig. 5C, 5D). The banded zones of staining

extended into the fish, appearing to mark skeletal muscles responsible for erecting or depressing the fins. The caudal fin also displayed a similar banded pattern of staining (Fig. 5F).

Besides the fins, *FoxP2* marked other non-neural structures in late larval embryos. Recalling intense expression at the fins, the skeletal muscle along the flanks (the myotomes) also showed *FoxP2* expression, though more diffusely than at the fins (Fig. 5B). There was also some maintenance of transcript production around the gut, particularly in its most torsional areas immediately posterior to the hindbrain, though discrete organs within the gut's length have not yet become elaborated (Fig. 5A). Meanwhile, the swimbladder has become a fully discernible organ. Expression is light but visible, particularly in the epithelium in the coronal view (Fig. 5A, 5B).

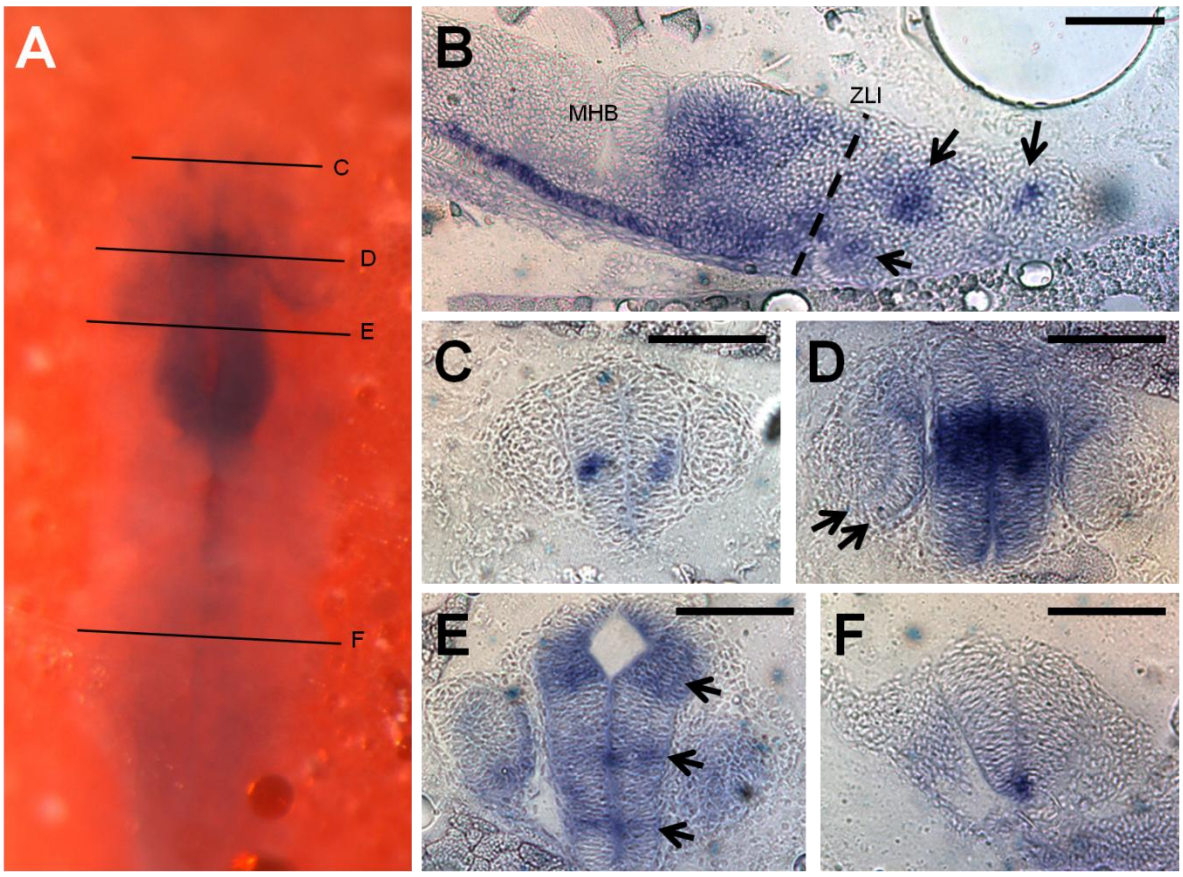


Figure 2: *FoxP2* expression in *M. zebra* at 3 days post-fertilization. (A) Whole-mount in dorsal view. (B) Sagittal section. Arrowheads point to prominent foci of expression within the prosencephalon. (D) Arrowheads point to two layers of expression within the prospective eyes. (E) Arrowheads point to three paired bands of expression in the developing midbrain. (F) Section further posterior with the neural tube in cross-section. *FoxP2* is expressed at the base of the neural tube. Scalebars are 100 μm. MHB, mid-hindbrain boundary; ZLI, zona limitans intrathalamica.

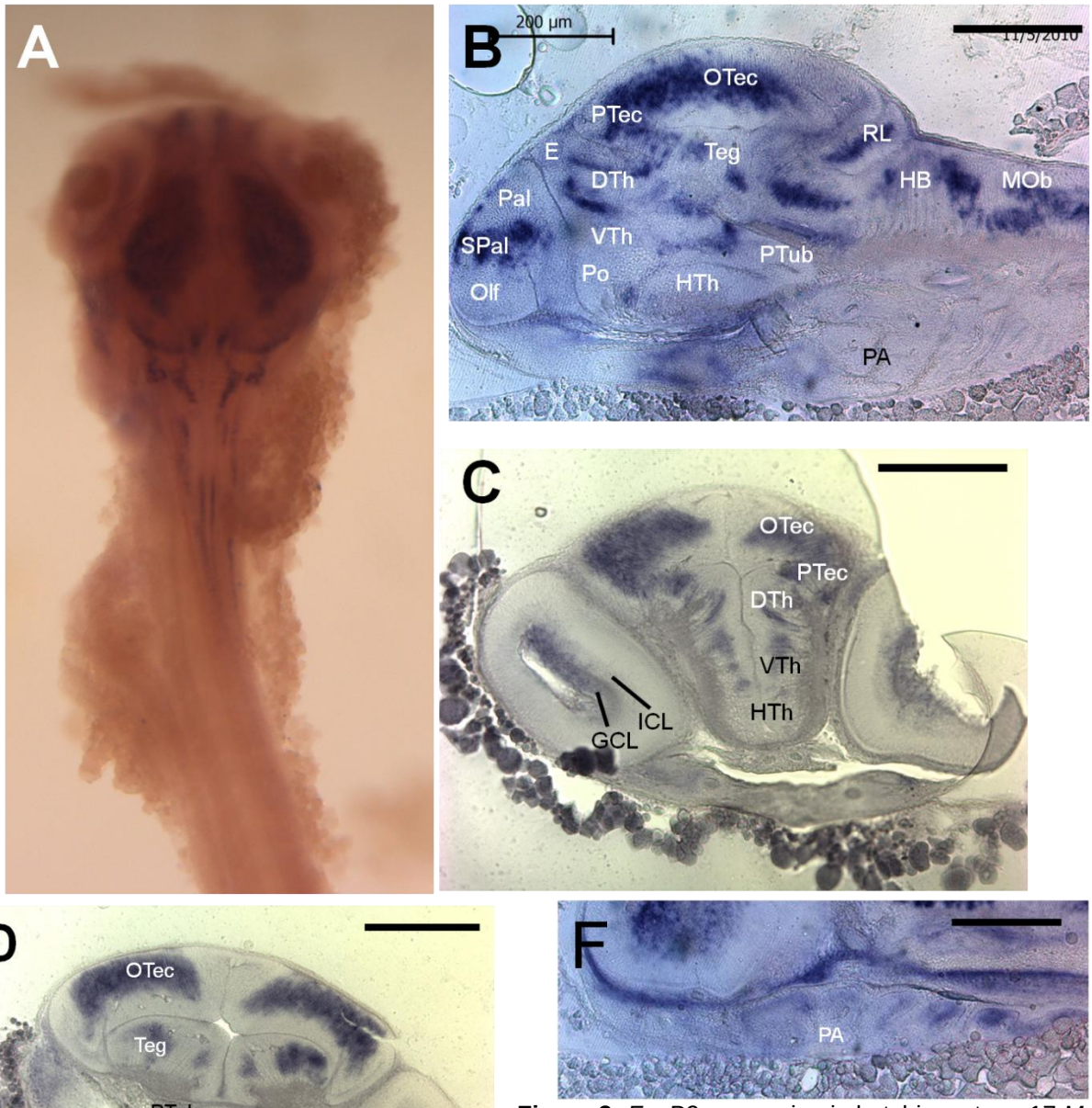


Figure 3: *FoxP2* expression in hatching, stage 17 *M. zebra* embryos (5 days post-fertilization). (A) Whole-mount visualization. (B) Sagittal section near the midline. (C-E) Coronal sections. Expression is visible in certain parts of the forebrain, most notably the subpallium, parts of the dorsal and ventral thalamus, and the pretectum. In the midbrain, the optic tectum and tegmentum both express *FoxP2*. The rhombic lip and other parts of the hindbrain also show expression. (F) Pharyngeal arches show expression in 5 day embryos, illustrated here better in sagittal view than in coronal view. Abbreviations: OTec, Optic tectum; DA, Dorsal aorta; DTh, Dorsal thalamus; E, Epiphysis; G, Gut; HB, Hindbrain; HTh, Hypothalamus; MOb, Medulla oblongata; Olf, Olfactory bulb; Nc, Notochord; NT; Neural tube; PA, Pharyngeal arches; Pal, Pallium; Po, Preoptic region; PTec, Pretectum; PTub, Posterior tuberculum; RL, Rhombic lip; SPal, Subpallium; Teg, Tegmentum; VTh, Ventral thalamus. Scalebars are 200um.

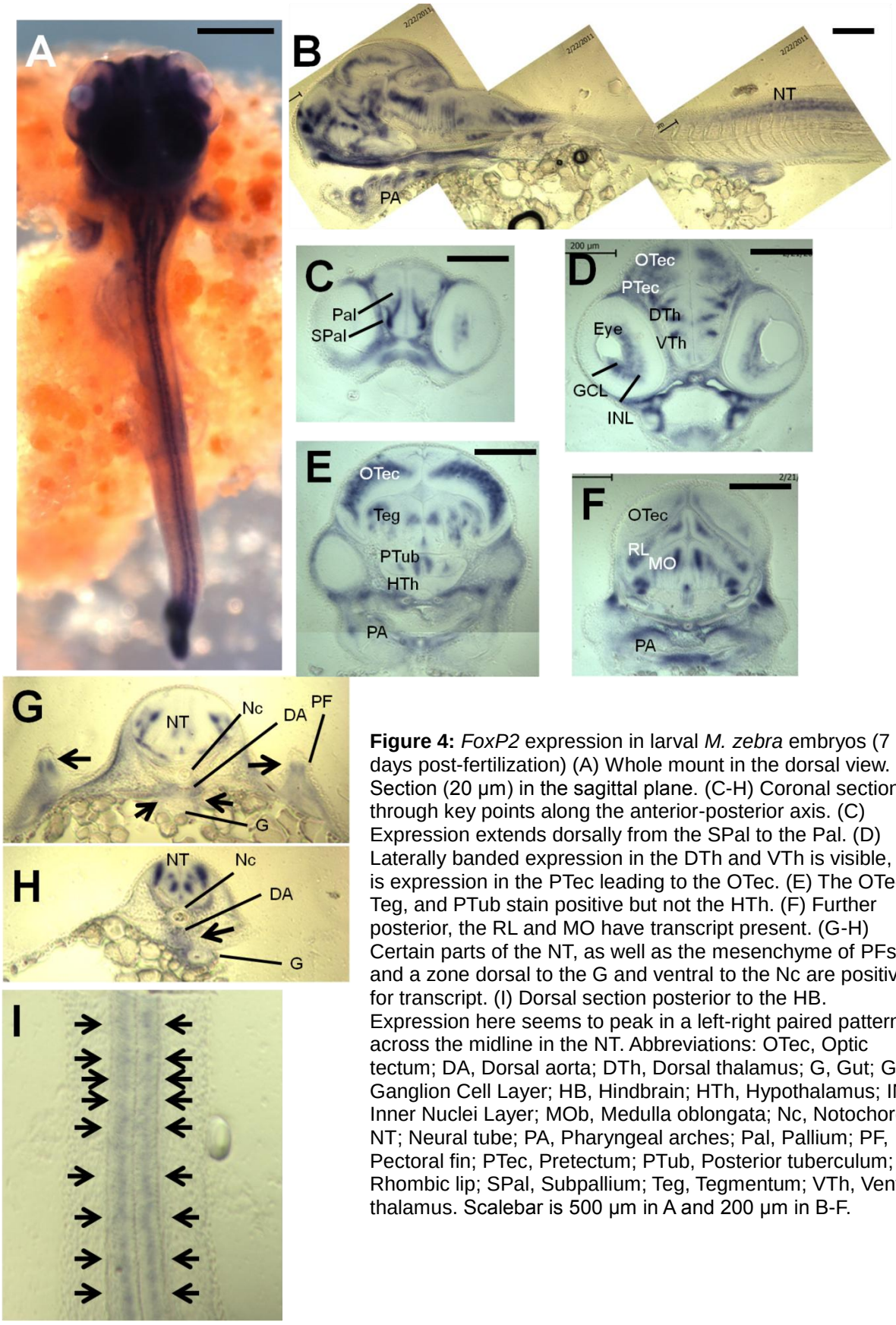


Figure 4: *FoxP2* expression in larval *M. zebra* embryos (7 days post-fertilization) (A) Whole mount in the dorsal view. (B) Section (20 µm) in the sagittal plane. (C-H) Coronal sections through key points along the anterior-posterior axis. (C) Expression extends dorsally from the SPal to the Pal. (D) Laterally banded expression in the DTh and VTh is visible, as is expression in the PTec leading to the OTec. (E) The OTec, Teg, and PTub stain positive but not the HTh. (F) Further posterior, the RL and MO have transcript present. (G-H) Certain parts of the NT, as well as the mesenchyme of PFs and a zone dorsal to the G and ventral to the Nc are positive for transcript. (I) Dorsal section posterior to the HB. Expression here seems to peak in a left-right paired pattern across the midline in the NT. Abbreviations: OTec, Optic tectum; DA, Dorsal aorta; DTh, Dorsal thalamus; G, Gut; GCL, Ganglion Cell Layer; HB, Hindbrain; HTh, Hypothalamus; INL, Inner Nuclei Layer; MOb, Medulla oblongata; Nc, Notochord; NT; Neural tube; PA, Pharyngeal arches; Pal, Pallium; PF, Pectoral fin; PTec, Pretectum; PTub, Posterior tuberculum; RL, Rhombic lip; SPal, Subpallium; Teg, Tegmentum; VTh, Ventral thalamus. Scalebar is 500 µm in A and 200 µm in B-F.

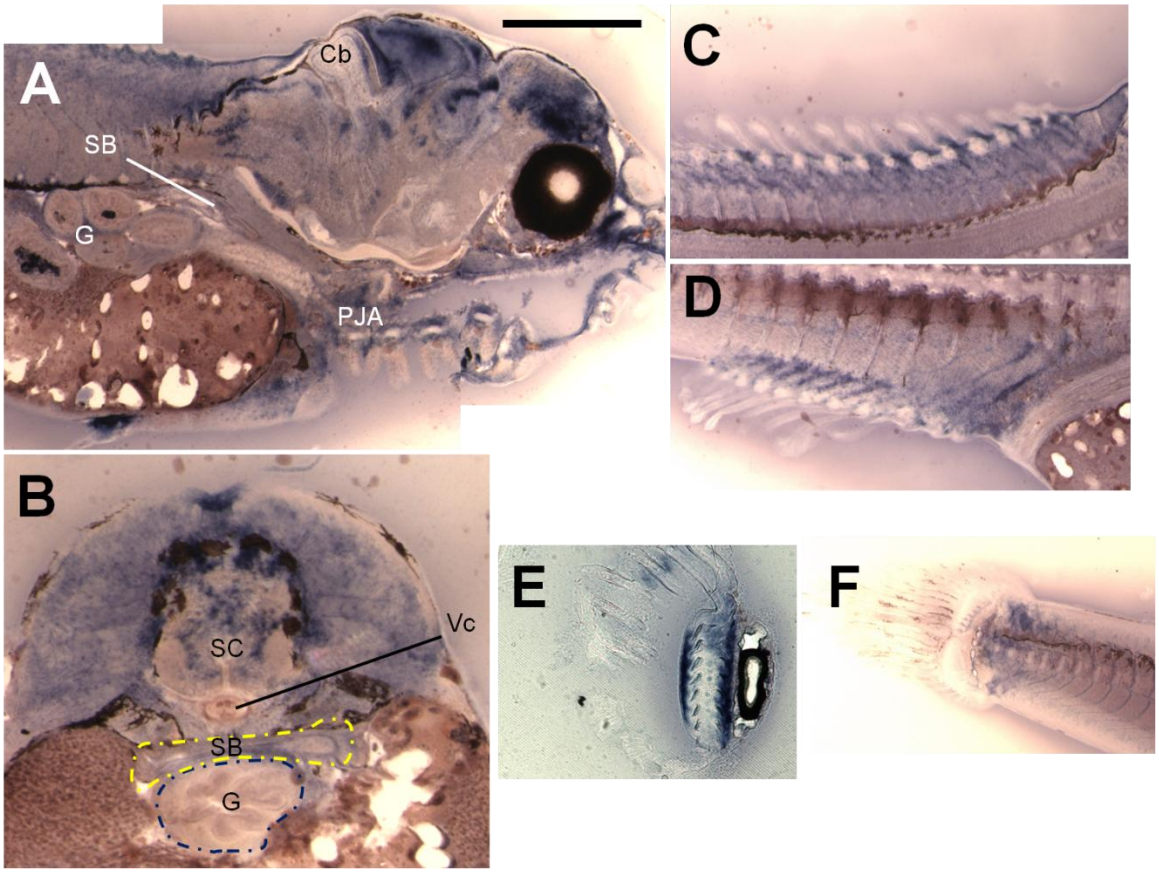


Figure 5: Hybridization in late larval (20 day) *M. conophorus* embryos. (A) Sagittal view. Swimbladder is well-defined but uninflated. Expression is faintly visible on the exterior of the G and in the SB. The oral jaw has extended and PJA has emerged from the pharyngeal arches. Expression is still visible in structures deriving from all seven of the pharyngeal arches. Brain expression has changed dramatically, still strongly expressed in the optic tectum but now limited to the anterior extent of the Cb (previously the rhombic lip). (B) Coronal view just posterior to the hindbrain. *FoxP2* expression is more evident in the SB in this view. NT displays complex expression pattern. Skeletal muscles along the flanks display transcription in a diffuse pattern. Additionally, the dorsal fin (C), anal fin (D), pectoral fins (E), and caudal fin (F) all show very active transcription, particularly in the erector or depressor muscles of the medial fins in C and D. Abbreviations: Cb, Cerebellum; G, Gut; PJA, Pharyngeal Jaw Apparatus; SB, Swimbladder; SC, Spinal cord; Vc, Vertebral column. Scalebar in A is 500 um.

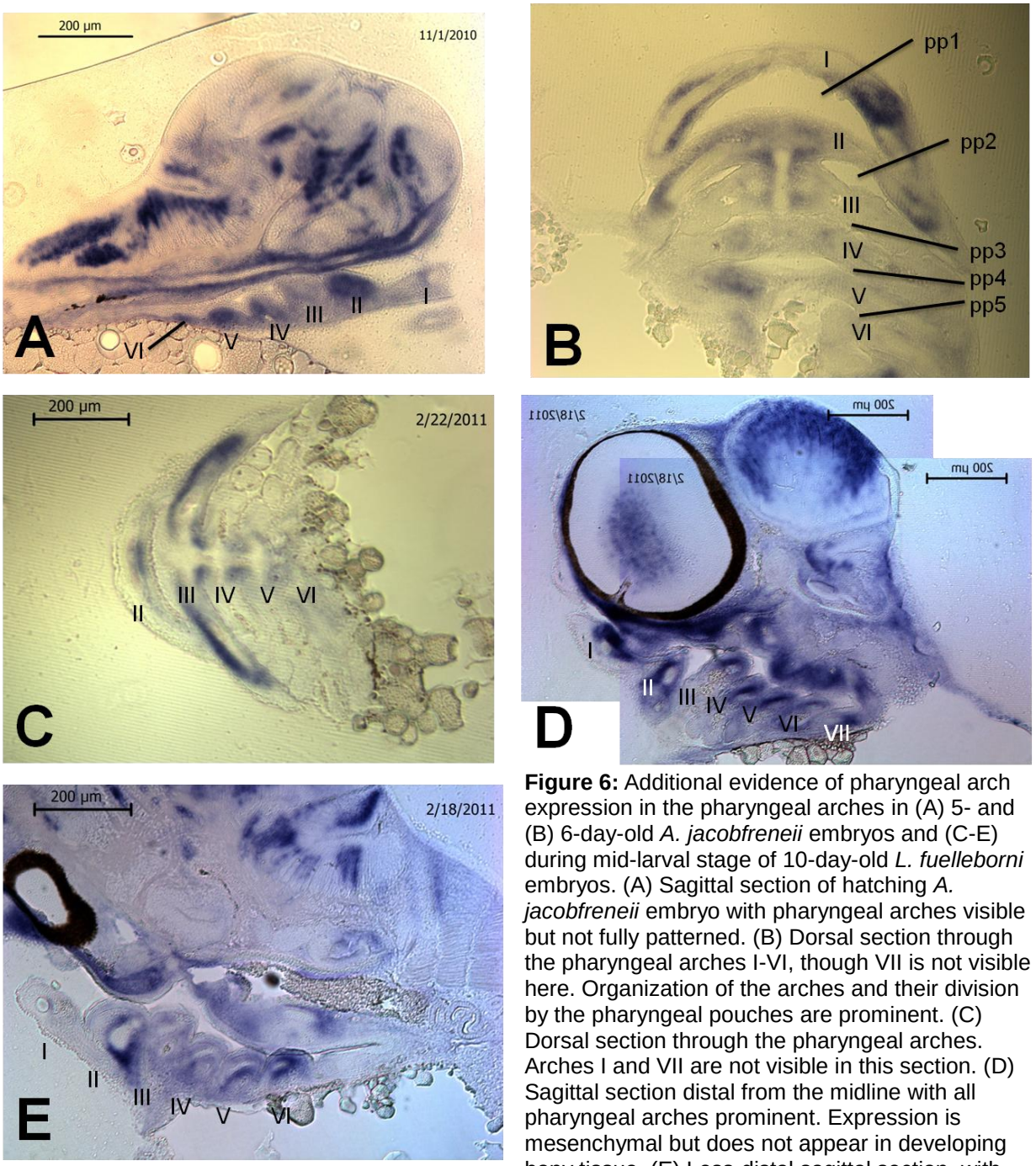


Figure 6: Additional evidence of pharyngeal arch expression in the pharyngeal arches in (A) 5- and (B) 6-day-old *A. jacobfreneii* embryos and (C-E) during mid-larval stage of 10-day-old *L. fuelleborni* embryos. (A) Sagittal section of hatching *A. jacobfreneii* embryo with pharyngeal arches visible but not fully patterned. (B) Dorsal section through the pharyngeal arches I-VI, though VII is not visible here. Organization of the arches and their division by the pharyngeal pouches are prominent. (C) Dorsal section through the pharyngeal arches. Arches I and VII are not visible in this section. (D) Sagittal section distal from the midline with all pharyngeal arches prominent. Expression is mesenchymal but does not appear in developing bony tissue. (E) Less-distal sagittal section, with Arch VII not visible. In all views, expression is extremely localized to discrete areas within all seven arches, interior to the ectoderm and marking the mesenchyme surrounding ossified tissue. I-VII, pharyngeal arch 1-7; pp#, pharyngeal pouch following PA#.

Genetic analysis of *FoxP2*

Summary of sequencing

The polymerase chain reaction was used to amplify 137 tandem overlapping amplicons along the *FoxP2* gene for each of seven cichlid species. Primers are listed in Supplementary Table S2. These reactions were sequenced as described, and assembled in Sequencher to build a consensus sequence. Due to spontaneous failures of amplification or sequencing for some reactions, data was sometimes absent for some species in a given reaction. Amplicons targeting difficult sequence such as low-complexity repetitive regions often failed sequencing. Nevertheless, a minimum of five species' sequences were present at nearly all points across the region. All successful sequencing reads were assembled in Sequencher and yielded a consensus sequence of 74,933 bases. Each species' assembly is described in more detail in Table 1.

Sequencing began approximately 7 kb upstream of the 5' start site of translation and terminated about 6 kb downstream of the 3' end site of translation. The length of transcribed sequence including introns but excluding UTRs was approximately 58kb. Coding sequence represented a small minority of that sequence, totaling 2,268 bases plus an Opal stop codon, while the remaining 56 kb were intronic.

Table 1: Summary of sequencing data obtained from each species to yield the consensus aligned draft sequence of 74933 bases. *Mbuna* lineage are highlighted in light gray; dark gray highlights non-*mbuna*.

Species	Usable bases	Gaps	Ns
A: *Copadichromis eucinostomus*	58710	3730	12493
B: *Cynotilapia afra*	71552	2660	721
C: *Protomelas taeniolatus*	71769	2579	585
D: *Labeotropheus fuelleborni*	71592	2620	721
E: *Metriaclima zebra*	63236	3035	8662
F: *Tramitichromis brevis*	71578	720	2635
G: *Mchenga conophorus*	73688	524	721

Annotation of the cichlid *FoxP2* coding sequence

Cichlid exons were annotated using a local alignment of the previously identified draft *FoxP2* transcript against the assembled genomic sequence. The draft transcript aligned to the genomic sequence in 16 fragments approximating exons. These loci were manually inspected for intron splice signals (GT/AG, GC/AC, etc.) and for adherence to the expected frame of translation (Fig. 7B). Subsequently they were lined end-to-end as a coding sequence (CDS), superseding the previous 1X draft transcript, and translated (Fig. 9). Therefore, we have identified 16 exons that together produce a mature *FoxP2* messenger RNA in cichlids (Fig. 7A).

Conservation of coding and non-coding sequence across divergent taxa

The draft consensus assembly (74.9 kb) was locally aligned with the homologous genomic sectors in the Medaka, Tetraodon, Fugu, Stickleback, Zebrafish, Human, Chimpanzee, and Mouse. All alignments displayed significant colinearity of their plus strands without prominent rearrangements (Fig. 8). Strong overall similarity was visible for all models, and acutely visible in the Stickleback, Fugu, Medaka, and Tetraodon models.

Portions of non-coding sequence including introns were strongly and consistently conserved across all fish and mammal models that were aligned. The fish models showed very large areas of conservation throughout the introns with cichlid genomic sequence, while the mammalian models showed smaller but consistent areas of homology (Fig. 10). Large portions of sequence 5’ of exon 1 and 3’ of exon 16 were also conserved, even in mammals (Fig. 10).

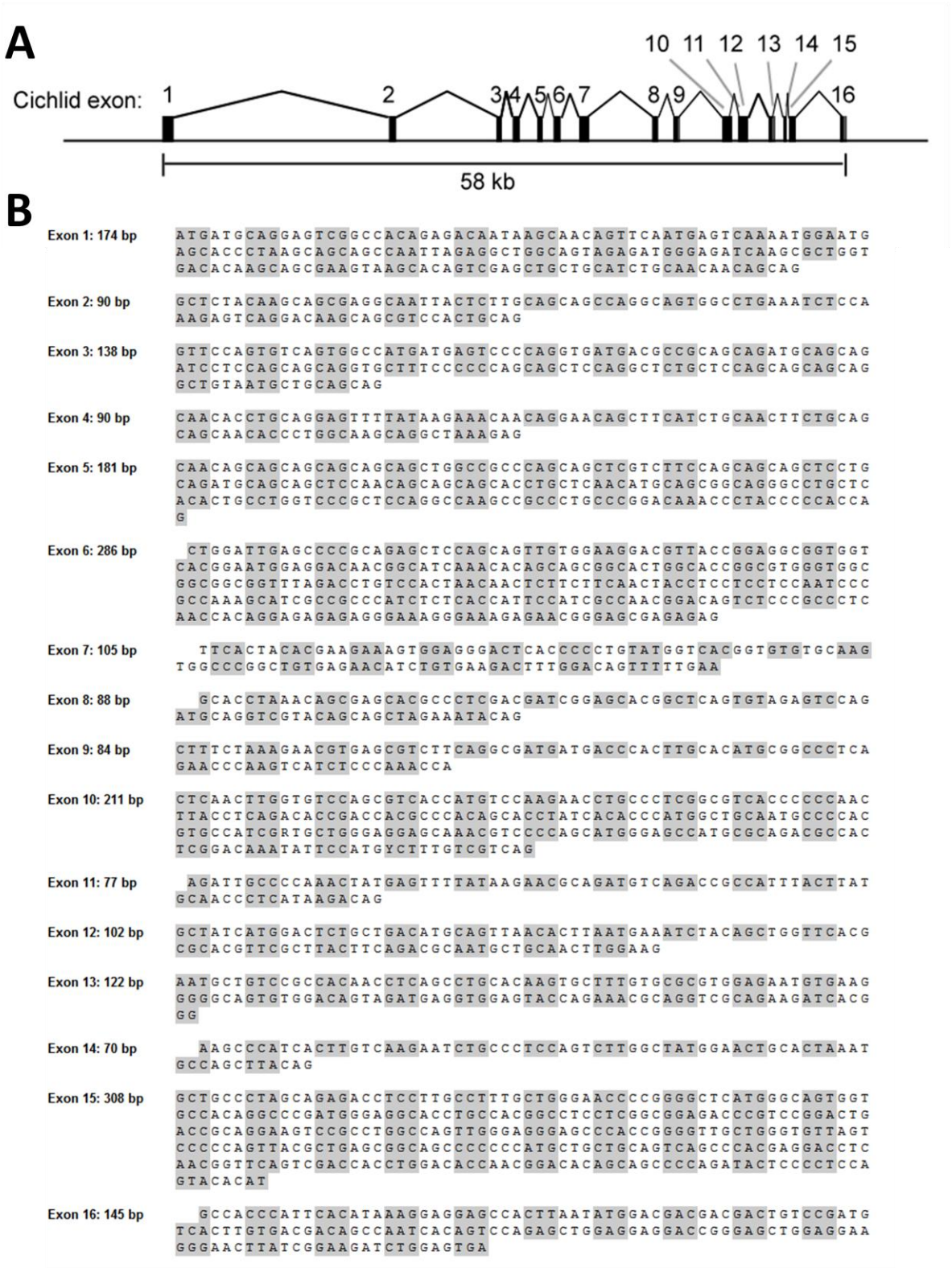


Figure 7: Exons within the cichlid *FoxP2* genomic sequence. (A) Gross illustration of transcript assembly. (B) The coding sequence of each exon, with codons marked. Exon 16 is terminated with TGA (UGA on the messenger RNA), the Opal translation stop signal.

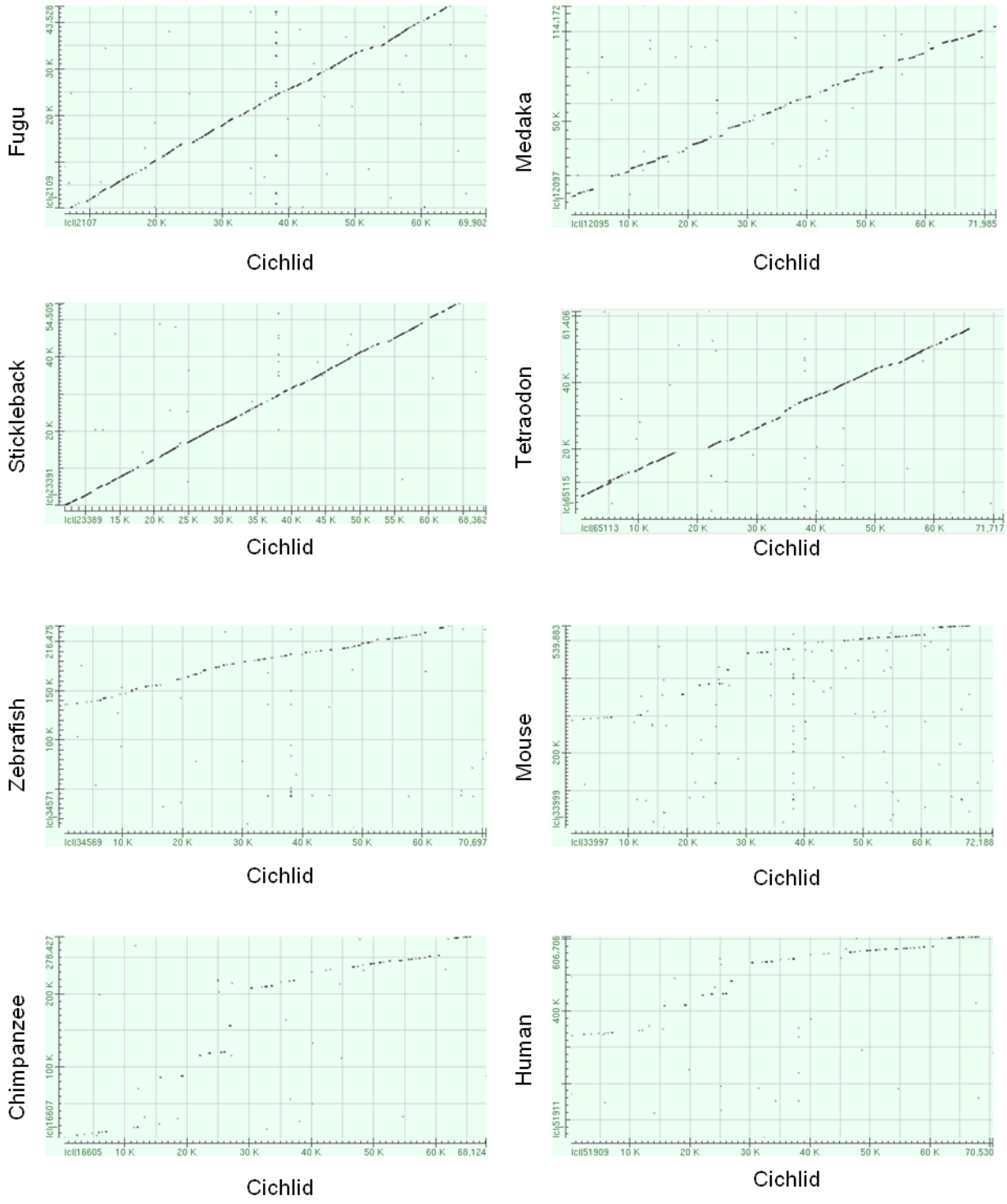


Figure 8: Dot matrix illustrations of local alignments of the cichlid draft assembly versus Fugu, Medaka, Stickleback, Tetraodon, Zebrafish, Mouse, Chimpanzee, and Human models. Similarity appears as black dots or lines, where longer lines signify more significant similarity implying homology. Plus-strand-to-plus-strand correspondence is indicated by a positive slope.

```
1   MMQESATETI SNSSMSQNGM STLSSSQLEA GSRDGRSSAG DTSSEVSTVE LLHLQQQQAL
61  QAARQLLLQQ PGSGLKSPKS QDKQRPLQVP VSVAMMSPQV MTPQQMQQIL QQQVLSPQQL
121 QALLQQQQAV MLQQQHLQEF YKKQQEQLHL QLLQQQHPGK QAKEQQQQQQ QLAAQQLVFQ
181 QQLLQMQQLQ QQQHLLNMQR QGLLTLPGPA PGQAALPGQT LPPPAGLSPA ELQQLWKDVT
241 GGGGHGMEDN GIKHSSGTGT GVGGGGGLDL STNNSSSTTS SSNPAKASPP ISHHSIANGQ
301 SPALNHRRER ERERERERER SLHEESGGTH PLYGHGVCKW PGCENICEDF GQFLKHLNSE
361 HALDDRSTAQ CRVQMQVVQQ LEIQLSKERE RLQAMMTHLH MRPSEPKSSP KPLNLVSSVT
421 MSKNLPSASP PNLPQTPTTP TAPITPMAAM PHVPSXLGGA NVPSMGAMRR RHSDKYSMXL
481 SSEIAPNYEF YKNADVRPPF TYATLIRQAI MDSADMQLTL NEIYSWFTRT FAYFRRNAAT
541 WKNAVRHNLS LHKCFVRVEN VKGAVWTVDE VEYQKRRSQK ITGSPSLVKN LPSSLGYGTA
601 LNASLQAALA ETSLPLLGTP GLMGSGATGP MGGTCHGLLG GDPSGLTAGS PPGQLGGSPP
661 GLLGVSPPVT LSGSPPMLLQ SAHEDLNGSV DHLDTNGHSS PRYSPPVHMP PIHIKEEPLN
721 MDDDDCPMSL VTTANHSPEL EEDRELEEGN LSEDLE*
```

Figure 9: The computational translation of consensus cichlid *FoxP2* coding sequence.

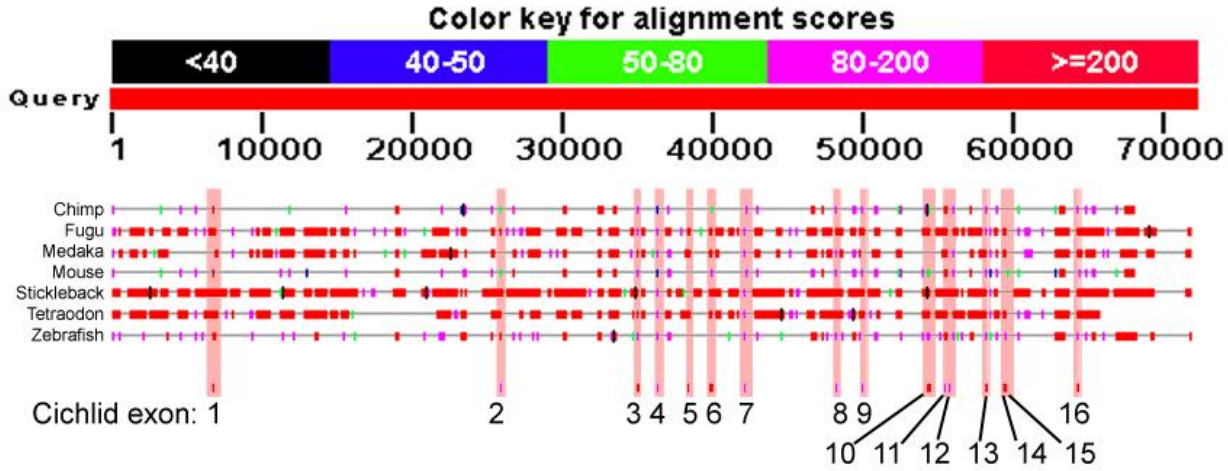
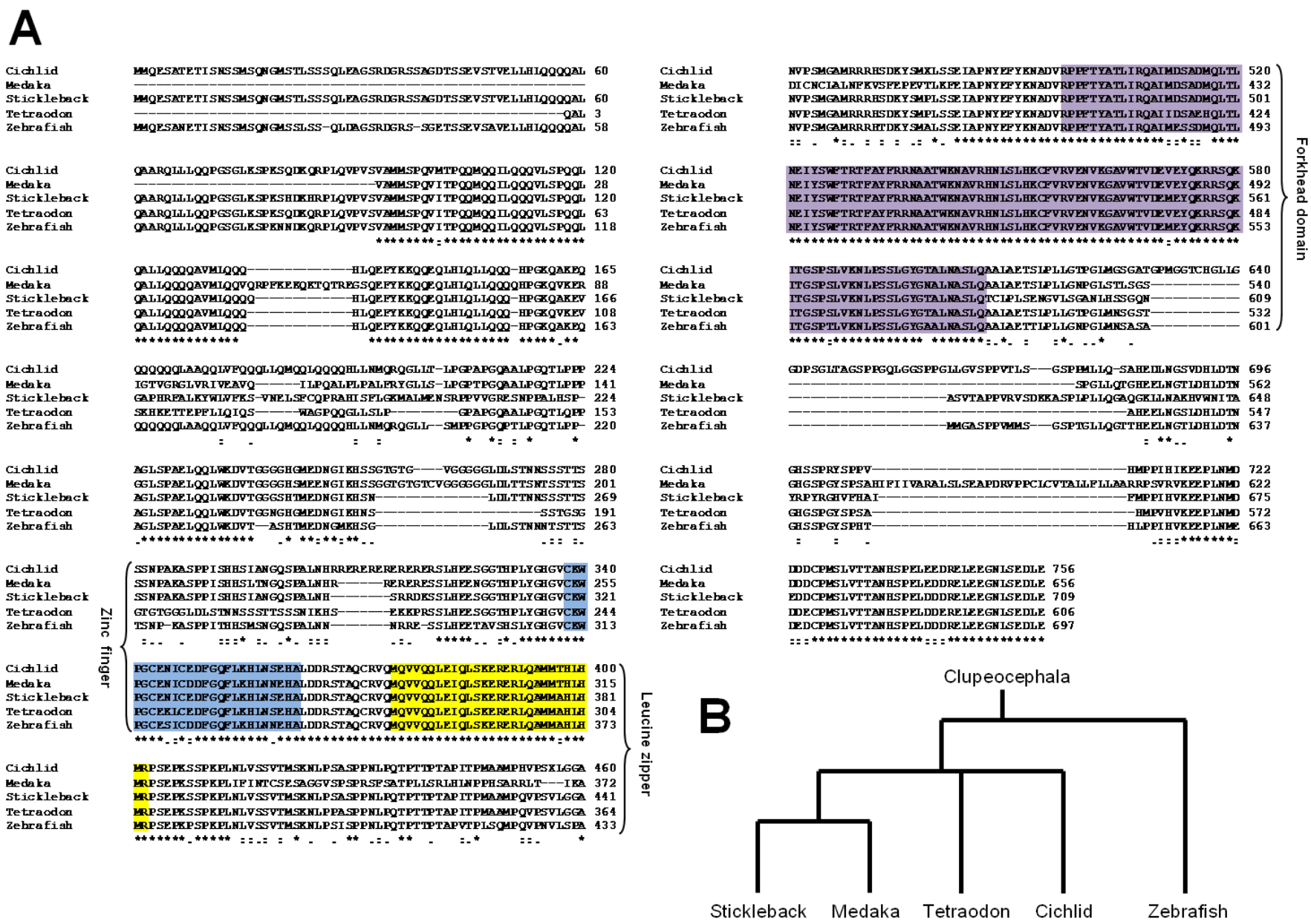


Figure 10: The genomic sequences of *FoxP2* across many lineages displays marked similarity, both in the arrangement of coding sequence as well as many discrete areas of non-coding sequence throughout the UTRs, introns, and non-transcribed sequence. Here, local alignments of genomic sequences illustrate mass conservation outside annotated cichlid exons.

Conservation of the *FoxP2* protein across divergent taxa

The *FoxP2* protein displayed visible sequence homology between orthologues in Cichlid, Medaka, Stickleback, Tetraodon, and Zebrafish (Fig. 11). Sequence conservation tended to adopt a mosaic pattern with less variability in some domains and greater variation in other areas. The Polyglutamine Domain (near the N-terminal area), the Zinc Finger Motif, the Leucine Zipper, the Forkhead Binding Domain, and the Acidic Domain (near the C-terminal area) all displayed very high sequence homology. Meanwhile, protein sequence between the Polyglutamine Domain and Zinc Finger Motif, as well as between the Forkhead Domain and Acidic domain, appeared somewhat less well-conserved.

Figure 11 (next page): (A) Alignment using ClustalW of the amino acid sequence of *FoxP2* in cichlids versus medaka, stickleback, tetraodon, and zebrafish. Exact matches are marked as "*" and similar amino acids are marked with ":" or ".". Selected functional domains are boxed and labeled. (B) Evolutionary tree of species shown based on taxonomy. The first exons of Medaka and Tetraodon were unavailable on Ensembl at the time of this writing, indicated by the absence of homologous protein sequence including a Start codon (Methionine).


A
Forkhead domain
Zinc finger
Leucine zipper
B
Clupeocephala
Stickleback
Medaka
Tetraodon
Cichlid
Zebrafish

Polymorphism in *FoxP2* coding sequence between different cichlid species

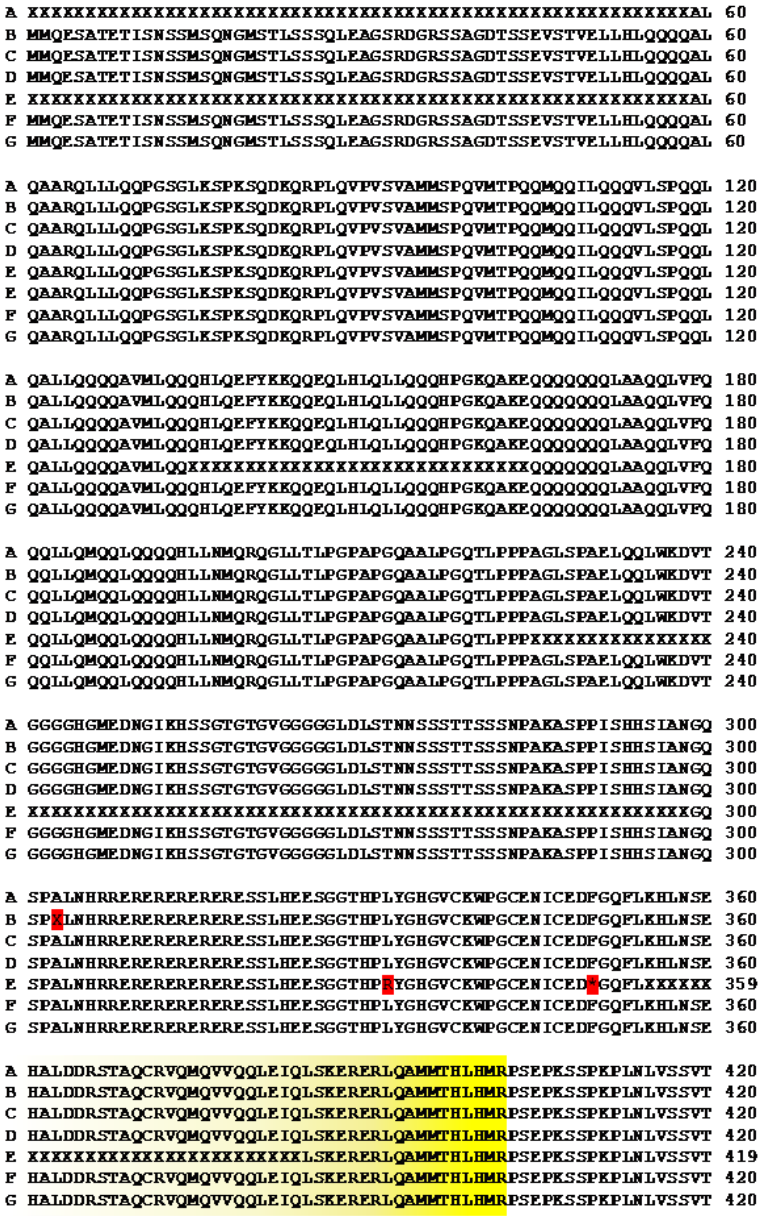
Despite the strong conservation of the *FoxP2* coding domain across wide taxa, sequencing across cichlid species revealed notable single nucleotide polymorphisms including polymorphisms causing amino acid changes (Fig. 12). Sequence data should still be considered 'draft' quality, and single outliers from the consensus should be treated with caution. However, two or more deviations from consensus provide stronger support for the existence of a SNP.

In particular, two amino acid polymorphisms were only separated by 22 amino acids and were flanked by the Leucine zipper domain (N') and the Forkhead domain (C'). These polymorphisms were in very strongly conserved amino acid positions (Fig. 13). At the first polymorphic position, Valine was present in all outgroup fish and mammals given. In cichlids however, three species maintained the Valine while four others displayed a novel Methionine. At the second position, Proline was conserved with the exception of Zebrafish. Again in cichlids, four species maintained the conserved Proline, and the other three had a Serine.

At least within our data set, these two amino acid polymorphisms were perfectly linked such that one predicts the other. A Valine was always followed by a Serine, while a Methionine always predicted a Proline. This pattern segregated according to cichlid lineage, where *Mbuna* (rock-dwelling) cichlids had a ValSer allele, and Non-*mbuna* had MetPro. Neither set of cichlid alleles (ValSer or MetPro) included the ancestral allele ValPro.

Figure 12 (next page): Multi-sequence alignment of the *FoxP2* protein reveals polymorphism. (A) Draft-quality alignment of computational translations of *FoxP2* CDS in 7 species. Two strongly supported amino acid polymorphisms are in the block of amino acids from position 421 to 480. These polymorphisms are framed by the Leucine zipper (yellow) and Forkhead domain (purple). Key: A = *Copadichromis eucinostomus*, B = *Cynotilapia afra*, C = *Protomelas taeniolatus*, D = *Labeotropheus fuelleborni*, E = *Metriaclima zebra*, F = *Tramitichromis brevis*, and G = *Mchenga conophorus*. (B) Summary of polymorphism locations and properties. Individual amino acid polymorphisms should not be taken as final until sequencing quality is improved further.

A



B

Position	Polymorphism	Frequency	Position	Polymorphism	Frequency
302	A/X (heterozygote or ambiguity)	7/1	456	V/M	4/4
332	L/R	7/1	479	S/P	4/4
350	F/Stop	7/1	550	S/X (heterozygote or ambiguity)	7/1

Figure 13: ClustalW protein alignment of translated cichlid exon 10 against homologous amino acid sequence in Fugu, Stickleback, Tetraodon, Zebrafish, Chimp, and Mouse. Cichlids with a *Mbuna* lineage are highlighted in light gray; dark gray highlights non-mbuna cichlids. Polymorphisms of interest are highlighted and compared to homologous positions in non-cichlids. The displayed amino acid polymorphisms appear exactly linked, where Val is always followed by Ser (blue letters), and Met is always followed by Pro (red letters). Underlining indicates deviation from evolutionarily conserved amino acid: Met is a deviation from Val, and Ser is a deviation from Pro. The polymorphisms look fixed between mbuna and non-mbuna populations.

K_A/K_S *p*-val	Ce		Ca		Pt		Lf		Mz		Tb	
Ce												
Ca	0.82	0.32										
Pt	-	-	50	0.2								
Lf	0.82	0.32	-	-	50	0.2						
Mz	50	0.32	50	0.37	50	0.33	50	0.37				
Tb	0	0.06	50	0.2	-	-	50	0.16	50	0.32		
Mc	0	0.06	50	0.16	-	-	50	0.2	50	0.32	-	-

Table 2: Pairwise K_A/K_S calculations using the Goldman-Yang method for all combinations of the seven cichlid sequences considered. Calculated ratios greater than 1 are given in green, and calculated ratios less than 1 are given in purple. Dashes indicate that the K_A/K_S ratio was undefined. Ce = *Copadichromis eucinostomus*, Ca = *Cynotilapia afra*, Pt = *Protomelas taeniolatus*, Lf = *Labeotropheus fuelleborni*, Mz = *Metriaclima zebra*, Tb = *Tramitichromis brevis*, and Mc = *Mchenga conophorus*. *Mbuna* lineage are highlighted in light gray; dark gray highlights non-*mbuna*.

Estimating selection through pairwise calculations of K_A/K_S

An estimator of selection, K_A/K_S, is roughly defined by the proportion of non-synonymous polymorphisms to synonymous polymorphisms. A ratio significantly greater than one indicates positive selection, and ratios significantly less than one indicate stabilizing selection. Selection for the whole *FoxP2* coding region was estimated using the Goldman-Yang method of K_A/K_S calculation in the KaKs_Calculator package (Goldman and Yang, 1994; Zhang et al, 2006).

This calculation was performed for each pairwise combination of 7 cichlid sequences considered (Table 2). Wide deviations were found for K_A/K_S values, highly contingent on the pairing of compared cichlid sequences. Many pairings such as B-C had ratios much greater than 1, while others such as D-E had ratios much less than 1. No cases revealed significant *p*-values. On a statistical hypothesis that there is either positive or negative selection, these data did not reject the null hypothesis of neutral selection.

Polymorphisms in non-coding sequence between cichlids, especially in conserved regions

Given the time allowed for cichlid species to radiate, we expected roughly a 0.2% incidence of polymorphisms in any given area of the genome (Loh et al, 2008). Over about 70kb, this rate predicted approximately 140 SNPs throughout *FoxP2* in Lake Malawi cichlids.

A total of 650 putative SNPs were found in the genomic region of *FoxP2* when considering all seven species of interest (species listed in Table 1). Of these, 637 were in non-coding regions. Many at non-coding areas (377) were in regions that were conserved in at least 1 other fish (any combination of Fugu, Stickleback, Tetraodon, Medaka, and/or Zebrafish). If we restricted consideration to areas where there were at least 5 cichlid base calls (i.e., fewer than 3

gaps or ambiguities) present, we were still left with 371 putative SNPs. As a collective, this data will be useful for estimating signs of selection like linkage disequilibrium, but is still being processed and filtered at the time of this writing.

SNPs in the most strongly conserved non-coding regions of *FoxP2* are potentially more interesting for individual inspection. Conservation against neutral genetic drift implies functional necessity for evolutionary fitness. SNPs in these regions are therefore more interesting candidates for additional screening and potential functional studies. By limiting consideration of SNPs to areas which are present in at least 3 fish and where at least 5 cichlid sequences are present, we can create a subset holding the most interesting SNPs for further analysis. After filtering as specified, there were 200 such putative SNPs.

The overall frequency of SNPs obtained in this analysis was noticeably higher than we expected. Again, given the length of genomic sequence and the time of divergence between species, we estimated there should be approximately 140 SNPs total in our data set. Sequencing read quality at this point is still a concern, even though some filtering of low-quality base calls was performed in Sequencher. Low-quality reads lead to a lower rate of accurate base calls representative of the true sequence, and lower accuracy produces disagreements which are not true SNPs. This issue prevents more detailed SNP analysis at this time, but shortly we will be re-sequencing low-quality areas and improving filtering techniques to reduce the incidence of non-representative disagreements.

DISCUSSION

Characterization of *FoxP2* expression in cichlids

FoxP2 in the nervous system

FoxP2 has prolific and complex patterns of expression within the embryonic brain. It is expressed in structures derived from all basic vesicular divisions of the brain, the prosencephalon (forebrain), mesencephalon (midbrain), and rhombencephalon (hindbrain).

In teleost fish, the prosencephalon everts into paired spherical structures, rather than undergoing significant infolding such as that in developing amniotes (Salas et al, 2003). Nevertheless, the fish forebrain is strongly homologous to the amniotic forebrain in both anatomy and function. The prosencephalon divides into the telencephalon and diencephalon, and the diencephalon is further divided by the zona limitans intrathalamica (ZLI).

The telencephalon forms the pallium and subpallium, which together are involved in space cognition and perception of the environment (Salas et al, 2003). The pallium provides several other sensory- and motor-related areas. *FoxP2* is strongly expressed in a complex and elegant spatiotemporal pattern in the telencephalon, implying a central role in the development of telencephalon-derived structures. In the prospective telencephalon, *FoxP2* is already being expressed in three concentrated areas (Fig. 2B). Later, small foci of expression are visible in the pallium, and extremely vigorous expression is visible in the subpallium (Fig. 3B). The olfactory bulb also forms from the telencephalon, but no *FoxP2* expression is found here at the timepoints considered (Fig. 3B).

The diencephalon forms the dorsal and ventral thalamus, preoptic tectum, hypothalamus, and other regions of the brain. The ZLI divides the dorsal and ventral thalamus and is also a primary signaling boundary for early brain development. *FoxP2* appears to define a discrete

population of cells in the ventral thalamus and thin bands of cells in the dorsal thalamus (Fig. 3C). Additionally, a small area of expression is visible on the posterior fringe of the preoptic region immediately anterior to the hypothalamus. However, the hypothalamus, responsible for autonomic nervous system control and hormonal signaling, does not express *FoxP2*.

On the other hand, the pretectum, which receives visual input from the retinal ganglion cell layer and relays the information to sensory processing centers such as the optic tectum, strongly expresses *FoxP2*. This expression domain is contiguous with the optic tectum (Fig. 3C). With relevancy to this system, the eyes display discrete layered patterns of *FoxP2* expression corresponding to specific cell populations: the ganglion cell layer and the inner nuclei layer. Expression is highest in the ganglion cell layer but still prominent in the inner nuclei layer. A picture emerges of *FoxP2* playing a major role in the development of visual pathways: the ganglion cell layer, the inner nuclei layer, the pretectum, and the optic tectum all express this transcription factor.

The mesencephalon is the embryonic origin of two prominent regions, the optic tectum and the tegmentum. The optic tectum, as with other brain structures, is strongly conserved across all vertebrates (Salas et al, 2003). It connects broadly with motor and sensory regions throughout the brain, contributes to awareness of orientation within the environment, and also directly helps coordinate and program muscle movements. The tegmentum has varied functions including roles in homeostasis and reflexes, but certain areas of the tegmentum also play a role in motor control and coordination (Kashin et al, 1974). *FoxP2* is expressed throughout development of the mesencephalon, emerging much earlier than its division into the tectum and tegmentum (Fig. 2B) and continuing throughout the development of the tectum and tegmentum as discrete regions (Fig. 3B, 3C, 3D). *FoxP2* is expressed throughout the optic tectum, but appears limited to

selective areas of the tegmentum (Fig. 3D). These data allow the possibility that *FoxP2* is present in motor control-related structures throughout the optic tectum and tegmentum, and not present in homeostatic or reflexive areas of the tegmentum.

The rhombencephalon, or hind brain, divides into a segmented pattern from rhombomeres and subsequently forms the metencephalon from anterior rhombomeres and myencephalon from posterior rhombomeres. The developing metencephalon bulges into a rhombic lip which will form the cerebellum, classically involved in learned motor responses and also recently implicated in learned emotional responses of teleost fishes (Rodrigues et al, 2005). The myencephalon forms the medulla oblongata which functions in many autonomic systems like heart rate and blood pressure control, enables temperature and pain perception, and also plays a role in coordination and some reflexes. *FoxP2* expression is found widely throughout the core of the rhombic lip but is biased towards some areas of the medulla oblongata, perhaps mirroring this region's varied function (Fig. 3B).

FoxP2 expression remains evident in the central nervous system posterior to the brain, extending throughout the spinal cord to its caudal-most point (Fig. 4A, 4B, 4H, 4I). Expression is continuous throughout the spinal cord, but seems to cluster in higher levels in paired formations throughout the cord's length (Fig. 4I). These may correspond to symmetrically paired ganglia which will relay stimuli to or from the peripheral nervous system to the CNS.

The pharyngeal arches and developing pharyngeal jaw

The pharyngeal or branchial arches are complex and highly ordered developmental intermediates which are critical for proper formation of many adult structures. In teleost fish, a total of seven pharyngeal arches (PA) contribute to the oral jaw (PA1), hyoid (PA2), and gills and

pharyngeal jaw apparatus (PA3-7). During pharyngula stage, PA1 arises individually as a mass of mesenchyme and PA2-7 similarly arise in a temporal pattern reflecting their position along the anterior-posterior axis. Though each arch maintains an individual identity which later contributes to differentiation of adult structures, all pharyngeal arches follow a common developmental process (Graham et al, 2005). *FoxP2* is expressed all pharyngeal arches (Fig. 5), and thus may contribute to this common process.

Endoderm, mesoderm, and ectoderm all contribute significantly to the development of pharyngeal arches. Endoderm and ectoderm form discrete populations of epithelial cells, and mesoderm forms the core of each arch. Pharyngeal pouches (pp), voids between each pharyngeal arch, are lined with endodermic epithelium and help direct arch patterning. Each pouch expresses *Bmp7* anteriorly, *Fgf8* posteriorly, and *Pax1* dorsally. Meanwhile, arch identity is maintained through spatially distinct expression certain genes within the endoderm. For example, the anterior endoderm of PA2 is marked by *Shh*. *HoxA* genes also contribute to PA endodermic identity, exhibiting canonical anterior-posterior nested expression patterns. *HoxA2* is expressed in PA3 and is limited anteriorly by pp2, but extends posteriorly into caudal PAs. *HoxA3* and *HoxA4* are limited in ranges more posterior to *HoxA2*. In this way, PA2 may be marked with *Shh*, PA3 with *HoxA2*, PA4 with *HoxA2* and *HoxA3*, and so on (Graham et al, 2005).

One of the most prominent and well-studied genes involved in pharyngeal arch development is the transcription factor *Tbx1*. *Tbx1* is required for early formation and outgrowth of all arches, and is particularly strongly expressed in the posterior arches. Its importance in human pharyngeal development is underlined in the symptoms of DiGeorge syndrome, which causes widespread problems including cleft palate, hearing loss, and velopharyngeal inadequacy. The syndrome is caused by a deletion in a chromosomal sector which includes *Tbx1* (Wurdak et

al, 2006). Intriguingly, *Tbx1* has at least one known binding motif in an enhancer responsive to *Fox*-family transcription factors (Yamagishi et al, 2002). This site agrees with the general consensus site described by Vernes et al (2007), and by definition is competent to bind *FoxP2*. Considering our data regarding *FoxP2* expression (Fig. 6), we have reason to suspect *Tbx1* may be a target of *FoxP2* regulation in the pharyngeal arches. In summary, *FoxP2* could conceivably mediate activation of *Tbx1* by *Shh* in the pharyngeal arches.

Though amniotes including humans possess 5 pharyngeal arches rather than 7, the basic mechanisms and ultimate functions of pharyngeal arch development are strictly conserved. It is likely that *FoxP2* is expressed in the pharyngeal arches of amniotes, and may indeed play a role in the development of its derived structures such as the oral jaw, the pharynx, the ear, and the larynx. Put simply, many structures derived from pharyngeal arches are related to vocalization, and *FoxP2* may help guide their development.

The foregut and swimbladder

The swimbladder is homologous in developmental origin to the tetrapod lung. Like the lung, the swimbladder derives from endoderm, budding from the foregut posterior to the pharynx. Three tissue layers emerge: the epithelium (facing the lumen), mesenchyme, and mesothelium (encasing the other layers). Wnt signaling is necessary for the early specification and proliferation of all three tissue layers, and is also sufficient to reprogram the specification of some gut tissue towards lung fate (Yin et al, 2011). In the mouse, *Wnt2* and *Wnt2b* are necessary for early lung progenitor specification, and no lung tissues will form in their combined absence. Later, Wnt signaling is still necessary for lung development. Selective loss of β-catenin in the epithelium or mesenchyme reduces proliferation of cells specifically in the layer of loss.

Of course, Wnts do not make a lung or swimbladder alone in a vacuum. Wnt signaling is modified and integrated through a variety of mechanisms. *Dkk1* antagonizes Wnt signaling, Frizzled receptors bind Wnts, and the transcription factors *Lef1* and *Tcf3* mediate downstream responses to Wnt signals (Yin et al, 2011; Hikasa et al, 2010; Boras-Granic et al, 2006). *Lef1*'s role as a downstream mediator of Wnt signaling is particularly interesting, because *Lef1* may indirectly affect Hh signaling (Boras-Granic et al, 2006). Though this proposed relationship of *Lef1* and Hh is unclear, Hh signaling is necessary for swimbladder and lung development (Winata et al, 2009).

As discussed previously, *FoxP2* is expressed along with *FoxP1* in the developing lung. *FoxP2* deletions in the mouse lung cause substantially reduced lung size and airway formation (Shu et al, 1997), and *FoxP2* has several enhancer elements which are responsive to *Lef1*. Possibly, *Lef1* regulates *FoxP2* expression in the lung in response to Wnt signaling. It is also tempting to say that *FoxP2* could provide a direct regulatory link between *Lef1* and a Hh family member, but *Shh* levels are unaffected by *FoxP2* deletions (Shu et al, 1997). Nevertheless, *Shh*, *Ihh*, and *Ptc1* are reduced when the Wnt antagonist *Dkk1* is overexpressed, indicating that Hh members do respond to Wnt signaling (Yin et al, 2011). Perhaps *FoxP2* mediates Wnt signaling through *Lef1* and acts downstream on *Ihh* or *Ptc1* rather than *Shh*. Or, perhaps it does not. To this author's knowledge, no studies have assayed *FoxP2* regulation of participants in the Hh pathway, other than *Shh*, in the swimbladder or lung. This hypothetical link might be interesting to investigate in the future.

In cichlid fishes, *FoxP2* is expressed in both the putative swimbladder and foregut at certain timepoints through their development (Fig. 4G, 4H, 5A, 5B). As expected, this expression recalls that of lung and esophagus development in mice (Shu et al, 1997). These data reaffirm

growing evidence that the teleost swimbladder and tetrapod lung are homologous.

Though I am confident in the identification of the cichlid swimbladder, I cannot claim without a doubt that what I have identified is indeed the swimbladder. In its identification, I used the neural tube, notochord, dorsal aorta, and gut as anatomical reference points. Ultimate proof must come from marking of the swimbladder with genes displaying swimbladder-specific expression, such as *Hp9*, *Fgf10a*, *Acta2*, *Sox2*, *Has2*, *Hprt1l*, or *Elovl1a* (Yin et al, 2011).

The pectoral fins, other fins, and other musculature

The induction of paired fins or limbs is an early event in development even though their morphogenesis only becomes visible later. During somitogenesis (starting roughly at 6-10 somites), retinoic acid-dependent pathways in the somatic mesoderm trigger inductive signaling cascades leading into the lateral plate mesoderm (Mercader, 2007). *Tbx5* is necessary for pectoral fin induction, and becomes expressed in presumptive pectoral fin mesenchyme even before a fin bud is visible.

Mesenchymal cells positive for *Tbx5* migrate to an increasingly concentrated area as the fin bud begins to condense. Cells expressing *Tbx5* trigger *Fgf24* at the fin field, ensuring compaction completes effectively. *Tbx5* continues expression as bud outgrowth begins in a process dependent on complex Wnt and Fgf signaling. At this time, *Tbx5*-expressing cells in the mesenchyme secrete *Fgf10* which initiates Fgf signals in the forming apical ectodermal ridge (AER) (Mercader, 2007). Continued growth leads to the formation of the achinotrichia (fin rays). Structures recognizable as fins, but still morphologically immature, are visible by 7 days post-fertilization in cichlids (though after 5 days post-fertilization).

At this point, *FoxP2* is expressed at the base of the fins. This area of expression is bounded anterior by the leading edge of the fins, and to the posterior continues to the trailing edge of the fins (Fig. 4A). Lighter projections of expression are visible in a radial pattern extending from the base. This expression is limited to the mesenchyme (Fig. 4G). The complex patterns of Wnt signaling could conceivably indirectly regulate *FoxP2*, though the only known regulator of *FoxP2*, *Lef1*, was shown to be dispensable for limb morphogenesis (Mercader, 2007). It is difficult to suggest a role for *FoxP2* in patterning the pectoral fins, but this would be an interesting topic of further research. It is still possible *FoxP2* participates in Wnt signaling, either responding to factors other than *Lef1* or driving Wnt signaling. Some ChIP data suggest parts of the Wnt pathway are responsive to *FoxP2* (Vernes et al, 2007; Spiteri et al, 2007).

On the other hand, after patterning has completed and the fins fully resemble their adult forms, *FoxP2* expression appears much more clearly defined (Fig. 5E). The mesenchymal tissues on the anterior and posterior sides of the achinotrichia strongly express the gene. The bony tissues themselves do not display such clear expression. At this stage in development, *FoxP2* seems to be marking near myoblasts or myocytes. This aligns neatly with Vernes et al (2007) and Spiteri et al (2007) whom predict a possible role of *FoxP2* in axon guidance.

Unlike the pectoral fins, which derive from lateral plate mesoderm, the dorsal, ventral/anal, and caudal/tail fins originate from somatic mesoderm and neural crest. They also emerge much later than pectoral fins in cichlid development. Even so, as their morphology approaches an adult form, *FoxP2* is expressed in a similar pattern to that in pectoral fins. The gene is expressed in mesenchyme near bony tissue but not in bony tissue itself (Fig. 5C, 5D, 5F). A banded pattern of expression which alternates with bony fin rays is visible, particularly in the dorsal and anal fins. These expression zones align with the erector and depressor fin muscles

which drive the fin rays. Dorsal fin muscles are anchored at the pterygiophores, and anal fin muscles are similarly anchored. In all cases, *FoxP2* appears to fully mark the fine muscles responsible for controlling the fins at the fish midline. Once again these data agree with the predicted role of *FoxP2* in axon guidance (Vernes et al, 2007; Spiteri et al, 2007).

Further implying a role of *FoxP2* in skeletal muscle innervation, expression data places the gene in the major muscles responsible for motive force while swimming, the myotomes (Fig. 5B). Expression in these muscles is noticeably lighter than in the fin muscles, but is nevertheless present. However, *FoxP2* is not expressed in the somites, precursors to the myotomes (Fig. 4H). The temporal specificity of *FoxP2* here imply that it is expressed in muscular tissue during innervation by the peripheral nervous system, but not long beforehand. The spaciotemporal coincidence of *FoxP2* and its predicted role in axon guidance strongly suggest *FoxP2* plays a role in the innervation of skeletal muscle. Further, its relatively stronger expression in the fin muscles compared to the myotomes imply it is more concentrated in the muscle tissues needing fine motor control and elaborate innervation. At this point it is not possible to say specifically if *FoxP2* is helping guide motor axons, various stretch receptors, or both. This matter would be an interesting subject of future investigation.

Conservation and diversity in the cichlid *FoxP2* gene

Conservation of certain motifs in genomic *FoxP2* sequence

The conservation of *FoxP2* genomic sequence extends greatly beyond exons to a plethora of conserved non-coding moieties. Even between such divergent species as mice and Lake Malawi cichlids, over thirty non-coding areas within *FoxP2* are conserved. These facts raise the likelihood that *FoxP2* has many sites receptive to regulation through any conceivable mechanism, from transcriptional regulation and control of alternate splicing to regulation by small RNAs both before and after transcription.

The transcription factor *Lef1*, an effector of the Wnt pathway, has been shown to have some regulatory sites on *FoxP2*. A total of 6 predicted Tcf/Lef binding sites are present on *FoxP2* in both mouse and zebrafish, and so far 3 have been verified as *Lef1*-responsive *in vivo* (Bonkowsky et al, 2008). *Lef1* is an important regulator of *FoxP2*, particularly in the tectum and hindbrain, but it is not the only regulator. Indeed, *FoxP2* is not expressed in every location showing *Lef1* expression, nor is *FoxP2* dependent on *Lef1* in the telencephalon (Bonkowsky et al, 2008).

Further *in silico* analysis of the conserved sequences in the non-coding areas of *FoxP2* will reveal many more putative regulatory factor binding sites. These data, once obtained, will be interesting to consider, but must be screened and verified either *in vitro* or *in vivo*.

Diversity in the coding domain, including non-synonymous polymorphisms

At least two key amino acid polymorphisms in the cichlid *FoxP2* protein give significant potential of differential allele-dependent protein action. The two polymorphisms of most intense interest are separated by 22 amino acids and are less than 20 positions N' of the Forkhead

domain. The first polymorphism involves a Met deviation at a conserved Val position, and the second polymorphism witnesses the option of a Ser at a conserved Pro position.

From the cichlid species studied, the polymorphisms appear strongly linked such that the alleles are either MetPro or ValSer, with no other combinations observed. Exactly one amino acid out of the pair of polymorphic loci deviates from the ancestral position. In other words, the ancestral allele ValPro appears rare in Lake Malawi cichlids. Further, the allele present tends to segregate between *mbuna* and non-*mbuna* populations. In all *mbuna* fish considered, the allele ValSer is present. On the other hand, non-*mbuna* fish apparently tend to possess the MetPro allele.

These amino acid polymorphisms provide ample material to be an intensely ermunicient source of vocal and behavioral diversity in cichlid fishes. Most critically, the polymorphisms' proximity (in 1º structure) to the Forkhead DNA-binding domain and the Leucine zipper domain imply potential ability to affect the functions of those domains. Drawing parallels to the evolution of human language, human *FOXP2* differs from chimpanzee *FOXP2* at two amino acids. The divergent positions here are about 50 amino acids N' of the Zinc Finger domain (Enard et al, 2002). Just as the amino acids of humans differ from other mammals at strongly conserved positions, the polymorphisms of cichlid *FoxP2* are present in tightly conserved areas close to functional domains.

As of yet, no conclusive evidence of positive selection has been found in the coding sequence of cichlid *FoxP2*. Calculations of K_A/K_S often give values greater than 1 (Table 2), but none with statistical significance. As we improve sequence quality above draft level, we may find many base disagreements are sequencing artifacts rather than true SNPs, and we may find other SNPs not previously detected. The refinement of data may change the net proportion of

observed coding and non-coding changes, or it may not. Regardless, refinement of our data will strengthen this important estimation of selection.

It is possible, however, that even refined data will not yield statistically significant calculations of K_A/K_S. This calculation is dependent on many SNPs to formulate positive selection. However, as we have stated previously, given the time of radiation cichlids have experienced in the lake, we only expect a frequency of approximately 0.2% SNPs per base position. In a coding sequence of 2268 nucleotides we would expect about 4-5 SNPs total. We will keep this number in mind as we continue to refine our data set.

CONCLUSIONS

We have decisively shown that *FoxP2* is much more than simply a speech and language gene, or even a gene only involved in development of brain motor control regions. It is involved in the development of motor control areas, of course, but also in brain areas relevant to sensory perception and processing. Our expression data also firmly support an equally interesting and largely unexplored side of *FoxP2*'s role in embryogenesis: the development of non-CNS structures which are still involved in motor force, vocalization, hearing, and vision. These data create a picture where *FoxP2* appears to be involved in the development of motor, sensory, and vocalization structures throughout the body. Its expression in development is not limited to one cell type or one cellular function, but is in fact expressed in cells deriving from all three embryonic layers.

Transcending tissue layer, cell type, or area or time of expression, the transcription factor might appear "messy" or poorly defined in purpose, acting in indecipherable or esoteric regulatory networks without uniting themes. Instead, I advocate that *FoxP2* plays a suite of functions that together integrate into logical, discrete set of capabilities characteristic of adult vertebrates.

Potential species-level differential regulation or downstream action of *FoxP2*

The sheer number and complexity of expression domains of *FoxP2* in cichlid fishes far exceeded our expectations. Our results to this point have been thrilling to parse and analyze, but they have also posed unforeseen difficulty due to their complexity. It is not yet possible, within our current dataset, to make statements about species-specific patterns of expression, nor is it possible to make claims regarding expression differences between species.

Differential changes of *FoxP2* may have widely pleiotropic effects throughout the body, inside and outside of the brain. Such distributed changes would, however, likely be small and difficult to detect with *in-situ* hybridization. Campbell et al (2009) have performed a comparison of *FoxP2* expression in adults of four species of mice which demonstrate diversity in their vocalization patterns, evoking similarity to cichlids. They found no major significant differences in expression pattern of any brain structure, though several minor areas of diversity were apparent. Such data reinforce the notion that studies of *FoxP2* expression diversity must have sufficient power to detect small differences in expression domain size or timing between species.

It is possible that upstream regulation of *FoxP2* operates in a species-dependent nature. Multiple polymorphisms exist in the non-coding areas of *FoxP2* including upstream of the translation start site, and many of these SNPs may affect the affinity of transcription factors to their binding sites. Effects may be small, perhaps undetectable on *in-situ* hybridizations, though may still be sufficient to explain species-specific patterns of vocalization established by Lobel (1998) and Amorim et al (2004). Species-specific differential regulation of *FoxP2* remains a distinct possibility to explain small vocal differences between species, though we cannot support this claim with expression data from *in-situ* hybridizations.

On the other hand, the downstream targets of *FoxP2* present exciting avenues for continued exploration of species diversity. Multiple amino acid polymorphisms in the protein's sequence, including some strongly segregating alternate alleles, raise the possibility that this protein operates differently species-to-species depending on the polymorphisms present. These amino acid changes, through shape changes or interactions throughout the protein's 3-dimensional structure, may affect stability, phosphorylation, protein-protein signaling, or DNA binding. These affects may manifest themselves in any area where *FoxP2* is expressed, including

the central nervous system, pharyngeal arches, swimbladder, sensory systems, foregut, fins, and skeletal muscles.

Genes responsive to *FoxP2* regulation may thus exhibit different transcription levels based on which *FoxP2* protein is present. Such regulatory action may be checked either *in vivo* or through such techniques as chromatin immunoprecipitation. At this present time, our lab is unable to check such effects, though this possibility should be considered in the future. As high throughput chromatin immunoprecipitation becomes possible for the still-underdeveloped cichlid model of developmental biology, such an experiment may prove fruitful.

Limitations of this study

Alternate transcripts were neither considered nor detected in assembling a sequence for a cichlid *FoxP2* transcript. Namely, our probe for *in-situ* hybridization was oriented towards our singly known transcript. In comparison, 4 total transcripts have been discovered for tetraodon, 7 for stickleback, and 21 transcripts have been described in humans. It is likely we have only constructed a sequence for the most prominent or actively expressed transcript of cichlid *FoxP2*s, as this transcript would be most easily amplified from total embryonic cDNA.

It is also possible that not all *FoxP2* exons in cichlids have been uncovered here, and may be present in more minor transcripts. We applied an effort to identify potential exons not included in our proposed transcript, searching for homology between other species' alternate transcripts and cichlid genomic sequence. Throughout the ~74.9kb genomic sequence, no sequences outside our transcript displayed significant homology to other known *FoxP2* proteins. Even so, sensitivity limitations may have prevented less well-conserved amino acid sequences from being detected.

An appeal for further investigation by fellow scientists

The research outlined in this report is still a work-in-progress. Additional *in-situ* hybridization experiments are continually being conducted to contribute to our growing library of expression data, and genetic data is subjected to unabated analysis and improvement of sequencing quality. We hope and expect that our twin surveys of expression and sequence will stand as comprehensive foundations to accelerate the growth in the molecular and functional understanding of *FoxP2*.

At this point, this author would like to speculate that when the common ancestor to *FoxP* subfamily genes was duplicated, *FoxP2* diverged from its paralogs to fill a functional niche in motor and muscle development. Perhaps *FoxP2*, freed from negative selective pressures thanks to its recent duplication, took on new functions and helped guide the co-evolution of a discrete set of motor, sensory, and communication functions, irrespective of cell type or origin. Over time, these functions would become entrenched, with *FoxP2* becoming strongly conserved via negative selection against deleterious changes in those functions. However, in cases where non-deleterious changes exist, the diversity could become a target of selection and source of divergence.

Diversity in motor, sensory, or communication functions, perhaps manifested through behavior, could accelerate evolution. *FOXP2* is believed to have been a primary contributor to the evolution of modern humans. Perhaps when diversity is present in social species, sexual selection on behavioral phenotypes results in fixation in *FoxP2* within breeding populations, contributing to mating barriers and driving species apart. If such a scenario is true, *FoxP2* could be an even more powerful contributor to evolution than previously thought, and its positive selection in humans would be far from unique.

Beyond evolutionary curiosity, though, further study into the gene *FoxP2* will likely have very real scientific and medical benefits. Its close relationship with Wnt signaling and its involvement with the development of complex integrated motor, sensory, and vocal systems make it an interesting target to investigate the genetic causes of perturbations of those functions. Patients with verbal dyspraxia suffer speech, balance, spatial orientation, and motor coordination problems. If the role of *FOXP2* in axon guidance to skeletal muscle fibers and/or stretch receptors is true, the observed coordination symptoms of patients may have an additional mechanistic basis outside the brain. Meanwhile, the expression of *FoxP2* in the pharyngeal arches and swim bladder suggest that the larynx and lungs of patients with R553H mutations may have problems originating in morphogenesis.

In conclusion, cichlids are interesting models for further study of the gene *FoxP2* and neatly supplement the strengths of more conventional models in developmental biology. Their diversity is both a microcosm of evolution and a natural experiment through diversity to determine the function of regulatory and coding sequences.

SUPPLEMENTARY TABLES

Table S1: Primers targeting the coding domain of cichlid *FoxP2* as designed from tilapia. Primers are given in 5'-3' order on the coding sequence they target. These are not given as primers designed or optimized specifically in pairs.

1E_L	GAGTCGGCCACAGAGACAAT	3H_R	CTGGAGGATCTGCTGCATCT
3H_L	TGTCAGTGGCCATGATGAGT	6H_1_R	CTGTGTTTGATGCCGTTGTC
6H_L	AGCAGTTGTGGAAGGACGTT	6H_2_R	GCGATGGAATGGTGAGAGAT
6H3_L	CCCTCAACCACAGGAGAGAG	8H_R	ACGACCTGCATCTGGACTCT
6H4_L	AGGGAAAGGGAAAGAGAACG	12H_R	CGTGATCTTCTGCGACCTG
7H1_L	GAGGGACTCACCCCCTGTAT	12H2_R	CGACCTGCGTTTCTGGTACT
8H_L	GCACCTAAACAGCGAGCAC	13H1_R	CAAGACTGGAGGGCAGATTC
12H_L	CACAACCTCAGTCTGCACAA	14H1_R	CAAAGGCAAGGAGGTCTCTG
		LE_1_R	TCTTCCGATAAGTTCCCTTCC
		LE_2_R	AAGTTCCCTTCCTCCAGCTC

Table S2: All primers used to amplify genomic sequence of *FoxP2* in cichlids. Primers were designed as pairs from tilapia as described in the Methods. Primer pairs which did not result in successful PCR amplification are omitted.

G001F	GAGCAGGTGAGCTTGGAGTC	G001R	CCTGTACTTGCCCCAATGAG
G002F	AATGAATGATGCCAGGATGC	G002R	CAGCACCTTCGTTTGCTAGA
F01F	TCACATGTGCAGGTAGACTGG	F01R	CCTCACTTCTGAAACGCGTAG
G011F	CGTCATGGGCACATTCCTTA	G011R	TGCCTCCCATAGTGATTTGA
H012F	TATGGGGTCAGTCGCTCATT	H012R	AGGCACTTGTCTTCCATGCT
G014F	ATAGCAGGCGCTCTCACACT	G014R	TTTCTGGGCTACATACTGCAAA
G015F	AGTGTCAAGAGGAAGCTGCTG	G015R	TGTCCATACGTTAAAAGTTCTCTGA
G016F	CAGCAGTGAAATGCCATGAA	G016R	TTGTCTGGATTTTCTGGGTTG
H017F	AGGTTAAGCCTTACAGCTACGC	H017R	GCCCTGGAACAAAGCTTATC
G018F	TGGCATTCTCGTAACTCGTG	G018R	GATGACGCTCTGGGTAGGC
G019F	CCCATACAAGAGGAGGCAGA	G019R	TGGAAGTGCAGAAACATTGTG
F02F	TCTTCATGTTTGGTTAGCGATG	F02R	TTCTCGCTGTTCTCCTTTTACC
G021F	GGACCAAGAGGAGTTTGGTG	G021R	ACTCCACAGTCCAAGCAACC
H022F	GTATGTGGCCACGCCATC	H022R	CCCACAAGCCTCTATCTGCT
F03F	CTTGCAGCACGATCACATTT	F03R	GCTTATCGGCCTTACAGAGG
G031F	CTGCATGGACTGTGTTCAGC	G031R	TTAGTGGCCCGACACTTCTT
G032F	AATAAGCACTGGGCAGGAAG	G032R	CCTGAGGCACAGAAACAGAA
G033F	GATGACTTTAATATTCAGAGCCAAAT	G033R	TGTGCTCGACAAAACTAGGG
F04F	TGTTCATAAACGCCATGCTG	F04R	ATCACACTGAGGCCCAACTC
G041F	GGGCAGGAGAAGTAGTGTGG	G041R	CGCACTAGCCTTGACACTTG
F05F	TTGAGGTTGTTGTGGCTTTG	F05R	GGGTTTAGCTGCTGTATTTTATGC
G051F	GAGGGATGAGGAAGGGAGAG	G051R	GCACATCATCTGCACTTTTCTC
G052F	ACCGGCAGCTTTTAATGAGA	G052R	GGCAGGTGAGGGGAAGTAGT
G053F	TCACACAGTTGTTTTGGTGGT	G053R	CGGGTGTTACGTGATTGACA
H053F	TGTTTGATGGCAGACTGGAA	H053R	AAATCAATGGTGCGTGAATG
G054F	TTCAGTCACTCAGGCAGTGG	G054R	AGCGATGTGTCCATTAAGAGC
F06F	TGGCTTATGGTGGTGACAGA	F06R	CACCCCACTTTTCCTTTCTC
G061F	TAATGCCCATGGCTCTGAAT	G061R	TGCAGCACATGTTTTAAGCAG
G062F	AATGACATACAGGGGCCAAG	G062R	CTCCCGTGCACTTTCACTC
G063F	ACTTGAGCGGGAGGAAGG	G063R	GGAAAATGTTGATTTGATGTGC
H063F	AAATGGCGGTCTGACATCTG	H063R	CGGGACGTCTTTTGATGTTT

G064F	GGGGCACATCAAATCAACAT	G064R	AACTACTCAGCGCAGCACCT
G065F	TGAGAATAAGCCACTTCATGTTAAA	G065R	TGCTCATTTCCTCAGTGCAT
F07F	ACGATCAGCCTCCGTGTTAT	F07R	TGACATATGCACCACCCAGT
G071F	ATCAAAGCAACTCCGTGGAC	G071R	AGCCTCTCACATGAGATTTGG
H071F	CCCAACAACACAACTGAAAGC	H071R	CCACCGGAGTGTCTCAAACT
F08F	TCGCTAATCGTCCAGCTTTT	F08R	AGGCGTCAGTTGTGTGTTTG
G081F	ACACTCCAAGTGCTCCCTGT	G081R	TCAGGGCTGAGAGTGAGTCC
G082F	ACATCCATCTACGGCATGG	G082R	TTAGCGTGTTTCTGCTGCTT
G083F	GGAGTTTGATTCCTGATGGA	G083R	GCTGATGACTGTTTGGCTCA
G084F	AAGCCACTGAGCCAAACAGT	G084R	GCAGACACTCAGGACCACCT
G085F	GAAGGTGCAGACAGGTCAGG	G085R	TACACTTGTGTTGCGTGTCC
G086F	AGACAGGACACGCAACACAA	G086R	CATGCTGGCTAGATAATGAGAGC
H086F	ACCATGTGAAAGCCTGAACC	H086R	GGGCCTTACAACCAAAGTCA
G087F	GCTTGTTGGCAGTGCTGAGT	G087R	GTTGATACTGTAATTATCATCCTCTGA
G088F	AAGAGAAGAGATGACAAGCCAAC	G088R	GCTCATTACCCAGCATCACA
F09F	GGACACAAAAGCTCCACCAT	F09R	TCTGCTGCTATCCCTCATCC
G091F	CAATTAGCCTGCGCACATT	G091R	GCACAGTAGGACATACAGCCATT
G092F	GCTTGCTGCATATGCTCATC	G092R	AGAGAAACTGCAGACTCCAATAA
G093F	AGCCAAGGTAGCGAAATCAA	G093R	CCCTCAGTAATGTACCTGCAAA
G094F	GCGGGTTACAGACACTAGGTT	G094R	AATAGCTCTTCAGTGCGCTTT
G095F	ACAAACACACCCACAAAGCA	G095R	CAGAACAAAGATGGCTGCTC
G096F	CCACTTTATGTAACCCCGTTT	G096R	AACAGTTTGCCTTGCAGCTT
G097F	TACAGCCATGCATCCTTTTG	G097R	GACAGAGGCAGGCAGTACATT
F10F	CACATGCAACTGGCTTTTCA	F10R	TGTTGCCAGGATGTTAGTGG
H101F	GTTTGTGCGAATTCACATGG	H101R	TCCCTCGAGGTAGTGCTGTT
G101F	AACCTGTAATCGGCAGTGCT	G101R	AGTCAGTTTTCGCTGCCATC
G102F	GCAGAGGGACAGCAACTGTA	G102R	TGCTGTCTAGGAAGGGTCGT
F11F	TGAGGACAAGGGGAGCAG	F11R	ATCTGGCCCATGCATTATTC
G111F	TGAGTCTGTACCAAGCATCTACAC	G111R	CCCATCTCTCACCATTCCAT
G112F	GGCGGATGGCATTTTAGTT	G112R	GTCGTTGTGCGTCAGAGCTA
G113F	TTTATTGAAGCCACGTGCAG	G113R	TGCCGCTGTGTCTCTCATAC
G114F	CATACCAATGAAGGGAAAGCA	G114R	CAGAGAGGGAGAGAGCGTGT
F12F	TGCTGTGACCTTAGCTACACCT	F12R	CCCCATTAAGCCTCAGTGAA
H121F	GAGAGATTTAAAGCTGCATGTGAA	H121R	CCACCACCACTCCATTATGC
G121F	CAGGTAATGATTGGCTGCAA	G121R	AAACACATGCAGCACCTCTG
G122F	GCTCGTCTACTGCTCCATGA	G122R	TGTGAAAGACTGCAAAGATCTGA
G123F	CTCCTGCAGGTGTTGCTGTA	G123R	CGATGTGGTCACATGCTTTC
H124F	CAGTGCTGCAGGGATAACAA	H124R	TGACGTACTCCTGCATGTCAC
F13F	GCACAGTAGAGCGTGAAAGC	F13R	TGCACCGTATGAACATTGTG
H131F	AGTCCCTGATGGATGCTTGT	H131R	ACCACACTTCCAGCAGATCC
G131F	ACCAGCCTAGGTGATGATGG	G131R	AGTTGCTTTCACACCGTTCC
F14F	GTCATGGGAGTTTGGGTTGT	F14R	TTATTGGCCCCTTGACACTC
H141F	ACGCTGCTTATGCCACTTTC	H141R	TTGAGCCAGCACCCAATTA
G142F	CAAATCCTGCCTCCATTGTT	G142R	TGCAAAGAGAGCTTCCTCATC
H142F	ATTAATTGGGTGCTGGCTCA	H142R	TTTATGGGTTGTCAGTGAAGAA
G143F	GCAACCAAGTTATGAGATGAGG	G143R	GCTGCAGTACTGGAAATGTATCCT
F15F	CAGGAGGGCTTGATTGACAT	F15R	TGGATGAGCTTGACCTTTCA
G151F	CTTGTCACACCTGCACCCTA	G151R	TGGGATGCTTAGGAGATTGAG
G152F	CATGTAGCCTCCCTACATTTCC	G152R	TCAGTCGGGAGTGAATTGTG
F16F	TGTAGCTGTGCGACATTGTG	F16R	AGGCGAGGTAGCTGAGTTTG
G161F	TGCATGTGTAAGGGCTGGTA	G161R	CAGGTCCACACATTTCACCA
G162F	CTTAAGCCTGGCGTCACAG	G162R	CTGCACAAACACAAGCACAG
G163F	TTGAGATGTCTCGTGGATGG	G163R	CCATAGATCACGGCATGTAA

G164F	CCTCCGCATCATCAACTTCT	G164R	GGCTCCCAGCTAATCCATTT
F17F	AGAATGCACTGAGCCCTGTT	F17R	CGTTAACCTCACCCGTTGAT
G171F	GCACACAGGGGTGTGACTC	G171R	CGTCACATAGACGCGAGAGA
G172F	AACCCACTATCCAGGCTGAC	G172R	TCTTCCCAGGCTCTACAAGC
G173F	ACGCTGCTTGTCCTGACTCT	G173R	TGCAAATGTGGCAGTAATTTG
F18F	CACAGAGCTGTGAGCAAGGA	F18R	TTTCTGCCAGGCTCTTTCAT
G181F	GCTTGTGTGCGCTGAATG	G181R	GGCAGCTGTTGTCATGGATA
H181F	CTGGGCTAGCAGCTTCACTC	H181R	ATTGATCGAGGGGGTTGTTT
G182F	CTCAGGAATCAGGGATGCTC	G182R	ATTCACCAATCATGGCGAGA
G183F	AATAACACCTGCAGCAAGCA	G183R	ACTTTCCTTGGCATTTGGTG
G184F	CACTTTGGCTGGCTCTGAGT	G184R	TGGCTGTAGCTTATTACCCAGTT
F19F	AATAGGCATGCAGAGATCACC	F19R	TAACATCTGTTCCGCACACC
G191F	CAGGCCTTGACAGGATGTTT	G191R	TGGAGTGTCAGCGCATATCT
G192F	AAGCTAGAGGGAGTCACATGC	G192R	TGTCTGTTCTGGACCCTCGT
H192F	AAGGGCTGCTGTTCAACACT	H192R	TATCGGCCATAGCTTCTGTG
G193F	CCAGCATACAGTTGTTCCACA	G193R	AAGCAGATTCTCCATGGACCT
F20F	ACAAACCTATGGGTGAAATCTTAG	F20R	CATCAGTGGGAGAAGTTTACACA
G201F	GGATTGTAACACACAAGCATCAA	G201R	AATGTTGAGGAGCAGCTATGC
G202F	TTCTGTGCAAATACAATGGTG	G202R	TTCTTCTCAAACCTCTAAGTGATCC
H202F	TGGCTGAAAGTTGAAGAGTGC	H202R	GCATCTTAAAGCCAGAACAGC
G203F	GGTCAGGAACAAATTAGAGGTCA	G203R	TTGTGCACACGCATCTTATG
G204F	GCTTACTCTAGCTATTAAGGAAGCAA	G204R	TGCTCCTTTGCTTTAGTGGAA
G205F	ACCTCACCCACACATACGG	G205R	CAAGCACAAACCTTATGCATTT
G206F	ACTCAGCTGGGTCTGGAGAA	G206R	CTCAGAAGTTGAGGTCTCTACCA
G207F	CATTCCCTTCATAGCACTCCA	G207R	TGGTTTCACGCTTGAGTCTG
F21F	ATTCTGGCGAGGACTAACGA	F21R	GCAGTCAAGGTGGAGTGCTA
G210F	ATGGATTCCCAGCGTGACT	G210R	CCCGATGGTTTGTTGTGTATC
G211F	TGATGATACCTGGAATCTGCAC	G211R	AAACTGGCTCGGTGTTGAAT
G212F	TGTATTCAGGCTGTGTGTGAC	G212R	CTTCTTCCATGCTGCCTTCT
G213F	TTCCTATATAATTGCATGGTAATGTC	G213R	CCTCTGTCTCCCGTTCTCTG
G214F	CTTTGCAGCTTGACATTTCG	G214R	GGCCTAATCACCCAGACTGT
G215F	GCCAGGGCCAGTGTCTTAT	G215R	GCCAAGGTGCCATCTTACTT
G216F	TGGACCACTGCCTAGCTCTT	G216R	TTGTCATTGACCGGAGTTACC
G217F	CAACTAACAGGCTCATTATAGGG	G217R	CATTTGGAATGACATTGAGTGG
G218F	AGCTGCAGGTGGTATTTTGG	G218R	CACAGCACTGAGAGGTGGAA
F22F	TTAGCCGGACTGCTCATTTC	F22R	CCATGTGTGACCAGAACACC
H221F	GATGGGAATGGTCTCTTCCA	H221R	CACAGCAAGTTGAGGGTGAA
G222F	ATGGCACGCATCATTACAAA	G222R	GCATGACAGAGGGAGAGAGG
G223F	GAGTTAGGAGGCCCTCTGGT	G223R	CATCTGCAACAACAGCAGGT
G224F	ATTGTGCCTCCTGCCTATGT	G224R	GCAATGCAGGTGCTCTGAT
F23F	CCATTATCAAGCCCTCAGTCC	F23R	GGTTTGATTAAAGAAGAGAGTGGA
G231F	CACCAAATCACATGGTGCAG	G231R	GCAAGGTTCCTCCATAAGAGA
G232F	AAATGCACTCCGCAATCAC	G232R	CCCTGAAAACATCACCACCT
G233F	CAGCGCTCACTACTGATGGA	G233R	CACCATCTGATCTCTGCGAAT
G234F	TGGCCAGCTTAATCTCAACG	G234R	CTTGGCGGGTATTGTAGCTG
G235F	AAGCACGCATGTAAGACTGC	G235R	CCTGACGATGCAAGATTCAA
G236F	GCTGAAGGCACCAATGTTTC	G236R	CTGGGATCACAGATTTACCG
G237F	CAGCATGAGGATGTTCTCCA	G237R	TCGCTACAAGAAGCTGTAGGC
G238F	AGATTCTGTGCTTGAGTGAAATTA	G238R	CACCAGTGAAGCACTTTTGAA
G239F	GCTGCCGTGCCTTAGAATTA	G239R	TGACTGAATGTCTCCTATCACAGA
G240F	TGTCCAATGAGCCCCTACAC	G240R	CCGTGCTGCACTAACATTTC
G241F	TTCATTGTACACCAGCTATCCA	G241R	TCAGACCCATCCAACTGTCA
G242F	CGATCAGTCGGCTCTCATTT	G242R	AGGACTGCTGAGAGGGAATG

REFERENCES CITED

Amorim MCK, M. E., Stratoudakis Y, and Turner GF (2004) "Differences in sounds made by courting males of three closely related Lake Malawi cichlid species." Journal of Fish Biology **65**: 1358-1371.

Boë L-J, Heim J-L, Honda K, Maeda S, Badin P, and Abry C (2007) "The vocal tract of newborn humans and Neanderthals: Acoustic capabilities and consequences for the debate on the origin of language. A reply to Lieberman (2007a)." Journal of Phonetics **35**: 564-581.

Bonkowsky JL and Chien C-B (2005) "Molecular cloning and developmental expression of *foxP2* in zebrafish." Developmental Dynamics **234**(3): 740-746.

Bonkowsky JL, Wang X, Fujimoto E, Lee J, Chien C-B, and Dorsky RI (2008) "Domain-specific regulation of *foxP2* CNS expression by lef1." BMC Developmental Biology **8**(1): 103.

Boras-Granic K, Chang H, Grosschedl R, and Hamel PA (2006) "Lef1 is required for the transition of Wnt signaling from mesenchymal to epithelial cells in the mouse embryonic mammary gland." Dev. Biol. **295**(1): 219-231.

Bruce H and Margolis R (2002) "*FOXP2*: novel exons, splice variants, and CAG repeat length stability." Human Genetics **111**(2): 136-144.

Campbell P, Reep RL, Stoll ML, Ophir AG, and Phelps SM (2009) "Conservation and Diversity of Foxp2 Expression in Muroid Rodents: Functional Implications." J. Comp. Neurology **512**: 84-100.

Chamberlain NL, Driver ED, and Miesfeld RL (1994) "The length and location of CAG trinucleotide repeats in the androgen receptor N-terminal domain affect transactivation function." Nucleic Acids Research **22**(15): 3181-3186.

Cheng L, Chong M, Fan W, Guo X, Zhang W, Yang X, Liu F, Gui Y, and Lu D (2007) "Molecular cloning, characterization, and developmental expression of *foxp1* in zebrafish." Development Genes and Evolution **217**(10): 699-707.

Enard W, Przeworki M, Fisher SE, Lai CSL, Wiebe V, Kitano T, Monaco AP, and Pääbo S (2002) "Molecular evolution of *FOXP2*, a gene involved in speech and language." Nature **418**: 869-872.

French CA, Groszer M, Preece C, Coupe A-M, Rajewsky K, and Fisher SE (2007) "Generation of mice with a conditional Foxp2 null allele." Genesis **45**(7): 440-446.

Fryer G and Iles TD (1972) The cichlid fishes of the Great Lakes of Africa. Edinburg, Oliver and Boyd.

Fujita E, Tanabe Y, Shiota A, Ueda M, Suwa K, Momoi MY, and Momoi T (2008) "Ultrasonic vocalization impairment of Foxp2 (R552H) knockin mice related to speech-language disorder and abnormality of Purkinje cells." Proceedings of the National Academy of Sciences **105**(8): 3117-3122.

Glazko GV, Koonin EV, and Rogozin IB (2005) "Molecular dating: ape bones agree with chicken entrails." Trends in Genetics **21**(2): 89-92.

Goldman N and Yang Z (1994) "A codon-based model of nucleotide substitution for protein-coding DNA sequences." Mol. Biol. Evol. **11**: 725-736.

Graham A, Okabe M, and Quinlan R (2005) "The role of the endoderm in the development and evolution of the pharyngeal arches." J. Anat. **207**: 479-487.

Graham JB (1997) Air-Breathing Fishes: Evolution, Diversity, and Adaptation. San Diego, Academic Press.

Groszer M, Keays DA, Deacon RMJ, de Bono JP, Prasad-Mulcare S, Gaub S, Baum MG, French CA, Nicod J, Coventry JA, Enard W, Fray M, Brown SD, Nolan PM, Pääbo S, Channon KM, Costa RM, Eilers J, Ehret G, Rawlins JN, and Fisher SE (2008) "Impaired synaptic plasticity and motor learning in mice with a point mutation implicated in human speech deficits." Current Biology **18**: 354-362.

Hikasa H, Ezan J, Itoh K, Li X, Klymkowsky MW, and Sokol SY (2010) "Regulation of TCF3 by Wnt-dependent phosphorylation during vertebrate axis specification." Developmental Cell **19**(4): 521-532.

Kashin SM, Feldman AG, and Orlovsky GN (1974) "Locomotion of fish evoked by electrical stimulation of the brain." Brain Research **82**(1): 41-47.

Klein D, Ono H, Hurgm O, Vmek V, Goldschmidt T, and Klein J (1993) "Extensive MHC variability in cichlid fishes of Lake Malawi." Nature **364**: 330-334.

Kocher TD (2004) "Adaptive evolution and explosive speciation: the cichlid fish model." Nature Reviews Genetics **5**: 288-298.

Konopka G, Bomar JM, Winden K, Coppola G, Jonsson ZO, Gao F, Peng S, Preuss TM, Wohlschlegel JA, and Geschwind DH (2009) "Human-specific transcriptional regulation of CNS development genes by *FOXP2*." Nature **462**(7270): 213-217.

Lai CSL, Gerrelli D, Monaco AP, Fisher SE, and Copp AJ (2003) "*FOXP2* expression during brain development coincides with adult sites of pathology in a severe speech and language disorder." Brain **126**: 2455-2462.

Lanzing WJR (1974) "Sound production in the cichlid *Tilapia mossambica* Peters." Journal of Fish Biology **6**(4): 341-347.

Li S, Weidenfeld J, and Morrisey EE (2004) "Transcriptional and DNA Binding Activity of the Foxp1/2/4 Family Is Modulated by Heterotypic and Homotypic Protein Interactions." Molecular and Cellular Biology **24**(2): 809-822.

Lieberman P (2007) "The Evolution of Human Speech: Its Anatomical and Neural Bases." Current Anthropology **48**(1): 39-66.

Lobel PS (1998) "Possible species specific courtship sounds by two sympatric cichlid fishes in Lake Malawi, Africa." Environmental Biology of Fishes **52**: 443-452.

Lobel PS (2001) "Acoustic behavior of cichlid fishes." Journal of Aquariculture and Aquatic Sciences **9**: 89-108.

Loh Y-HE, Katz LS, Mims MC, Kocher TD, Yi SV, and Streelman JT (2008) "Comparative analysis reveals signatures of differentiation amid genomic polymorphism in Lake Malawi cichlids." Genome Biology **9**(7): R113.

Longrie N, Van Wassenbergh S, Vandewalle P, Mauguit Q, and Parmentier E (2009) "Potential mechanism of sound production in *Oreochromis niloticus* (Cichlidae)." Journal of Experimental Biology **212**(21): 3395-3402.

MacDermot KD, Bonora E, Sykes N, Coupe A-M, Lai CSL, Vernes SC, Vargha-Khadem F, McKenzie F, Smith RL, Monaco AP, and Fisher SE (2005) "Identification of *FOXP2* Truncation as a Novel Cause of Developmental Speech and Language Deficits." The American Journal of Human Genetics **76**: 1074-1080.

Mercader N (2007) "Early steps of paired fin development in zebrafish compared with tetrapod limb development." Develop. Growth Differ. **49**: 421-437.

Perry SF, Wilson RJ, Staus C, Harris MB, and Remmers JE (2001) "Which came first, the lung or the breath?" Comparative Biochemistry and Physiology Part A: Molecular & Integrative Physiology **129**: 37-47.

Ptak SE, Enard W, Wiebe V, Hellmann I, Krause J, Lachmann M, and Paabo S (2009) "Linkage Disequilibrium Extends Across Putative Selected Sites in *FOXP2*." Molecular Biology and Evolution **26**(10): 2181-2184.

Rice AN and Lobel PS (2002) "Enzyme activities of pharyngeal jaw musculature in the cichlid *Tramitichromis intermedius*: implications for sound production in cichlid fishes." The Journal of Experimental Biology **205**: 3519-3523.

Ripley JL and Lobel PS (2004) "Correlation of acoustic and visual signals in the cichlid fish, *Tramitichromis intermedius*." Environmental Biology of Fishes **71**: 389-394.

Rodrigues F, Duran E, Gomez A, Ocana FM, Alvarez E, Jimenez-Moya F, Broglio C, and Salas C (2005) "Cognitive and emotional functions of the teleost fish cerebellum." Brain Res. Bulletin **66**(4-6): 365-370.

Salas C, Broglio C, and Rodriguez F (2003) "Evolution of Forebrain and Spatial Cognition in Vertebrates: Conservation across Diversity." Brain Behav. Ecol. **62**: 72-82.

Schroeder DI and Myers RM (2008) "Multiple transcription start sites for *FOXP2* with varying cellular specificities." Gene **413**: 42-48.

Shah R, Medina-Martinez O, Chu L-F, Samaco RC, and Jamrich M (2006) "Expression of *FoxP2* during zebrafish development and in the adult brain." The International Journal of Developmental Biology **50**(4): 435-438.

Shu W, Lu MM, Zhang Y, Tucker PW, Zhou D, and Morrisey EE (2007) "*Foxp2* and *Foxp1* cooperatively regulate lung and esophagus development." Development **134**(10): 1991-2000.

Spiteri E, Konopka G, Coppola G, Bomar J, Oldham M, Ou J, Vernes S, Fisher S, Ren B, and Geschwind D (2007) "Identification of the Transcriptional Targets of *FOXP2*, a Gene Linked to Speech and Language, in Developing Human Brain." The American Journal of Human Genetics **81**(6): 1144-1157.

Stauffer JRJ, Bowers NJ, McKaye KR, and Kocher TD (1995) "Evolutionarily significant units among cichlid fishes: the role of behavioral studies." American Fisheries Society Symposium **17**: 227-244.

Sylvester JB, Rich CA, Loh Y-HE, Straaden MJ, Fraser GJ, and Streelman JT (2010) "Brian diversity evolves via differences in patterning." Proceedings of the National Academy of Sciences **107**(21): 9718-9723.

Takahashi K, Liu F, Hirokawa K, and Takahashi H (2003) "Expression of *Foxp2*, a Gene Involved in Speech and Language, in the Developing and Adult Striatum." Journal of Neuroscience Research **73**: 61-72.

Teramitsu I, Poopatanapong A, Torrisi S, White SA, and Tanimoto H (2010) "Striatal *FoxP2* Is Actively Regulated during Songbird Sensorimotor Learning." PLoS ONE **5**(1): e8548.

Turner GF, Seehausen O, Knight ME, Allender CJ, and Robinson RL (2001) "How many species of cichlid fishes are there in African lakes?" Molecular Ecology **10**: 793-806.

Van Der Sluijs I, Van Dooren TJM, Seehausen O, and Van Alphen JJM (2008) "A test of fitness consequences of hybridization in sibling species of Lake Victoria cichlid fish." Journal of Evolutionary Biology **21**(2): 480-491.

Vernes S, Spiteri E, Nicod J, Groszer M, Taylor J, Davies K, Geschwind D, and Fisher S (2007) "High-Throughput Analysis of Promoter Occupancy Reveals Direct Neural Targets of *FOXP2*, a Gene Mutated in Speech and Language Disorders." The American Journal of Human Genetics **81**(6): 1232-1250.

Wang B (2003) "Multiple Domains Define the Expression and Regulatory Properties of *Foxp1* Forkhead Transcriptional Repressors." Journal of Biological Chemistry **278**(27): 24259-24268.

Winata CL, Korzh S, Kondrychyn I, Zheng W, Korzh V, and Gong Z (2009) "Development of zebrafish swimbladder: The requirement of Hedgehog signaling in specification and organization of the three tissue layers." Developmental Biology **331**(2): 222-236.

Wurdak H, Ittner LM, and Sommer L (2006) "DiGeorge syndrome and pharyngeal apparatus development." BioEssays **28**: 1078-1086.

Yamagishi H, Maeda J, Hu T, et al. (2002) "Tbx1 is regulated by tissue-specific forkhead proteins through a common Sonic hedgehog-responsive enhancer." Genes & Dev. **17**: 269-281.

Yin A, Korzh S, Winata C, Korzh V, and Gong Z (2011) "Wnt Signaling is Required for Early Development of Zebrafish Swimbladder." PLoS One **6**(3).

Zhang Z, Li J, Zhao XQ, Wang J, Wong GK, and Yu J (2006) "KaKs_Calculator: Calculating Ka and Ks Through Model Selection and Model Averaging." Genomics, Proteomics, & Bioinformatics **4**(4).